[illegible]
[illegible] OOL,
[illegible]-5002

NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

S8244

THE ESTIMATION OF UNITED STATES ARMY REENLISTMENT RATES

by

Michael J. Streff

September 1989

Thesis Advisor: Laura D. Johnson

Approved for public release; distribution is unlimited

classified

rity classification of this page

REPORT DOCUMENTATION PAGE

Report Security Classification Unclassified	1b Restrictive Markings
Security Classification Authority	3 Distribution Availability of Report Approved for public release: distribution is unlimited.
Declassification Downgrading Schedule	
erforming Organization Report Number(s)	5 Monitoring Organization Report Number(s)
Name of Performing Organization val Postgraduate School	6b Office Symbol *(if applicable)* 55
7a Name of Monitoring Organization Naval Postgraduate School	
Address *(city, state, and ZIP code)* onterey, CA 93943-5000	7b Address *(city, state, and ZIP code)* Monterey, CA 93943-5000
Name of Funding Sponsoring Organization	8b Office Symbol *(if applicable)*
9 Procurement Instrument Identification Number	
Address *(city, state, and ZIP code)*	10 Source of Funding Numbers

Program Element No	Project No	Task No	Work Unit Accession No

Title *(include security classification)* THE ESTIMATION OF UNITED STATES ARMY REENLISTMENT RATES

Personal Author(s) Michael J. Streff

Type of Report aster's Thesis	13b Time Covered From To	14 Date of Report *(year, month, day)* September 1989	15 Page Count 108

Supplementary Notation The views expressed in this thesis are those of the author and do not reflect the official policy or poion of the Department of Defense or the U.S. Government.

Cosati Codes		
ld	Group	Subgroup

18 Subject Terms *(continue on reverse if necessary and identify by block number)*
SRB, reenlistment, retention, Logistic Regression, Hierarchical Clustering, Exploratory Data Analysis

Abstract *(continue on reverse if necessary and identify by block number)*

The U. S. Army uses cash selective reenlistment bonuses (SRB) to encourage soldiers in selected military occupation ecialities (MOS) to reenlist. Estimates of the reenlistment rate as a function of bonus level are needed for each MOS as out to a bonus allocation model. This thesis outlines and uses a new method for predicting the reenlistment rates as a nction of bonus level.

The approach involves partitioning the soldier population into cells with stable reenlistment rates using demographic riables. The cells are aggregated using clustering techniques to produce groups of cells which exhibit homogeneity of enlistment behavior. Regression models are developed for each group of cells. MOS reenlistment rates are determined as inear combination across cells. Cross-validation techniques are used to lend credibility to the predictive model.

The study points out the usefulness of identifying categories of soldiers who display unique reenlistment behavior. Inteation of this technique with existing econometric reenlistment models is recommended to further improve the predictive odel.

Distribution Availability of Abstract unclassified unlimited ☐ same as report ☐ DTIC users	21 Abstract Security Classification Unclassified	
Name of Responsible Individual ura D. Johnson	22b Telephone *(include Area code)* (408) 646-2569	22c Office Symbol 55Jo

FORM 1473,84 MAR

83 APR edition may be used until exhausted
All other editions are obsolete

security classification of this page

Unclassified

T245495

Approved for public release; distribution is unlimited.

The Estimation of United States Army
Reenlistment Rates

by

Michael J. Streff
Captain, United States Army
B.S., United States Military Academy, 1979

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
September 1989

ABSTRACT

The U. S. Army uses cash selective reenlistment bonuses (SRB) to encourage soldiers in selected military occupation specialities (MOS) to reenlist. Estimates of the reenlistment rate as a function of bonus level are needed for each MOS as input to a bonus allocation model. This thesis outlines and uses a new method for predicting the reenlistment rates as a function of bonus level.

The approach involves partitioning the soldier population into cells with stable reenlistment rates using demographic variables. The cells are aggregated using clustering techniques to produce groups of cells which exhibit homogeneity of reenlistment behavior. Regression models are developed for each group of cells. MOS reenlistment rates are determined as a linear combination across cells. Cross-validation techniques are used to lend credibility to the predictive model.

The study points out the usefulness of identifying categories of soldiers who display unique reenlistment behavior. Integration of this technique with existing econometric reenlistment models is recommended to further improve the predictive model.

c.1

TABLE OF CONTENTS

I. INTRODUCTION 1
  A. GENERAL 1
  B. BACKGROUND 2
  C. RESEARCH QUESTIONS 3
    1. MOS Grouping 4
    2. Variables to be Considered 6
    3. Summary of Research Questions 8
  D. SCOPE OF THESIS 8
  E. ORGANIZATION OF THESIS 8
  F. STATISTICAL PACKAGES 9

II. REVIEW OF THE LITERATURE 10
  A. GENERAL 10
  B. ARMY STUDIES 10
  C. ACOL STUDIES 11

III. DATA BASE 13
  A. GENERAL 13
    1. Source of Data 13
    2. Response Variable 14
    3. Explanatory Variables 14
    4. Survey Data 14
    5. Time Period Covered 14
    6. Size of Data Set 15
  B. CONCEPTUAL FRAMEWORK 15
    1. Initial Motivation for Military Service 15
    2. Success in the Service and Satisfaction with Military Life 16
    3. Evaluation of Potential in the Civilian Sector 17
    4. Reenlistment Policy Variables 19
  C. SIGNIFICANCE OF UNQUANTIFIABLE VARIABLES 19
  D. CLEANING THE DATA SET 20

IV. METHODOLOGY 23
A. GENERAL 23
B. MOTIVATION FOR THE METHODOLOGY 23
1. Problems With Current Solution 23
2. Non-homogenous MOS 24
3. Example of Methodology 29
4. Assumption of the Methodology 29
5. Motivation for Variable Reduction 30
C. METHODOLOGY 30

V. ZONE A ANALYSIS AND RESULTS 31
A. GENERAL 31
B. SELECTION OF INFLUENTIAL CATEGORICAL VARIABLES 31
1. Exploratory Data Analysis of Categorical Variables. 31
2. Exploratory Data Analysis Tools 32
3. Distribution of Variables 32
4. Univariate Analysis 35
5. Multivariate Analysis 38
6. Table Selection 39
7. Results of Exploratory Data Analysis 40
C. PARTITIONING OF THE POPULATION INTO HOMOGENEOUS CELLS 41
D. CELL REDUCTION 41
1. Cell Reduction Procedure 41
2. Cell Reduction Results 42
E. SELECTION OF INFLUENTIAL CONTINUOUS VARIABLES 43
1. Exploratory Data Analysis of Continuous Variables 43
2. Distribution of Individual Variables 43
3. Univariate Analysis 43
4. Bivariate and Multivariate Analysis 47
5. Results of Exploratory Data Analysis 47
F. ESTIMATION OF REENLISTMENT RATES 48
G. COMPUTATION OF MOS REENLISTMENT RATES 48
H. MODEL VALIDATION 49
I. MODEL PRECISION 49

VI. CONCLUSIONS 52
  A. FINDINGS 52
  B. RECOMMENDATIONS 54
  C. RECOMMENDATIONS FOR FURTHER STUDY 54

APPENDIX A. THE MILITARY REENLISTMENT BONUS MODEL 55
  A. GENERAL 55
  B. INPUTS 55
  C. OUTPUT 55
  D. OBJECTIVE FUNCTION 55
  E. SOLUTION METHODOLOGY 55
  F. MODEL USE 56

APPENDIX B. CALCULATION OF REENLISTMENT RATES 57
  A. GENERAL 57
  B. MOS MIGRATION 57
  C. EXTENSIONS 58
  D. REENLISTMENT ELIGIBILITY 58
  E. EARLY REENLISTMENTS 59

APPENDIX C. VARIABLES TO MEASURE INITIAL MOTIVATION FOR MILITARY SERVICE 60

APPENDIX D. VARIABLES TO MEASURE THE SOLDIERS SUCCESS IN THE SERVICE 63

APPENDIX E. VARIABLES TO MEASURE A SOLDIERS POTENTIAL IN THE CIVILIAN SECTOR 65

APPENDIX F. REENLISTMENT POLICY VARIABLES 68

APPENDIX G. MISSING DATA 71
  A. PURPOSE 71
  B. MISSING DATA AFTER CLEANING 71
  C. RANDOM MISSING DATA 71

APPENDIX H. LOG-LINEAR MODELS 74

APPENDIX I. LOGISTIC REGRESSION 75

APPENDIX J. CLUSTER ANALYSIS RESULTS 81

APPENDIX K. REGRESSION ANALYSIS RESULTS 87

LIST OF REFERENCES 90

INITIAL DISTRIBUTION LIST 96

LIST OF TABLES

Table 1. REENLISTMENT RATES BY CATEGORY, FOUR VARIABLES . . . 27
Table 2. REENLISTMENT RATES COMPARISONS 29
Table 3. MEASUREMENT SCALE AND RANGES FOR CATEGORICAL VARIABLES 33
Table 4. REMAINING CATEGORICAL VARIABLES 39
Table 5. ASSOCIATIONS WITH COMPOUND VARIABLES 40
Table 6. RANGES, MEANS AND STANDARD DEVIATIONS FOR CONTINUOUS VARIABLES 44
Table 7. REENLISTMENT RATES FOR MOS 11B 49
Table 8. RESULTS OF MODEL VALIDATION 50
Table 9. MISSING DATA FOR CATEGORICAL VARIABLES 72
Table 10. CLUSTER RESULTS BY ZONE 83
Table 11. CLUSTER RESULTS BY ZONE (CONTINUED) 84
Table 12. REGRESSION RESULTS BY ZONE 88
Table 13. REGRESSION RESULTS BY ZONE (CONTINUED) 89

LIST OF FIGURES

Figure 1. Sample of Input Required for Bonus Model (Hypothetical) 4
Figure 2. Yearly Reenlistment Rates for Ten MOS's Over Seven Years 7
Figure 3. Frequency Counts for the Variable Term of Enlistment, Uncleaned . . . 21
Figure 4. Frequency Counts for the Variable Term of Enlistment, Cleaned 21
Figure 5. Reenlistment Rates for MOS 11B, Zone A by Dependent Status 25
Figure 6. Reenlistment Rates for Differing MOS's by Dependent Status 26
Figure 7. Reenlistment Rates for MOS 11B, Zone A by Race 27
Figure 8. Racial Composition of Three MOS's 28
Figure 9. Frequency Counts For Selected MOS's 34
Figure 10. Reenlistment Rates for all MOS's, by Age at Enlistment 35
Figure 11. Reenlistment Rates by Mental Category and by Rank 37
Figure 12. Reenlistment Rates for Regions of the Country 38
Figure 13. Regression of Bonus Level vs Reenlistment Probability 45
Figure 14. Plot of Bonus Level vs Reenlistment Probability 46
Figure 15. Breakdown of MOS 11B by Cell 51
Figure 16. Number of Observations With Missing Values 71
Figure 17. Reenlistment Rates for Observations With Missing Data 73
Figure 18. Linear Approximation to a Probability Function 76
Figure 19. Data Format for Logistic Regression 77
Figure 20. Number of Observations and Reenlistment Rates by Cell 85
Figure 21. Number of Observations and Reenlistment Rates by Cell 86

ACKNOWLEDGMENT

The author wishes to thank Dr. Bob Tinney of the Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia and Mr. Bruce McClellan of the Office of the Deputy Chief of Staff, U. S. Army, Washington, DC for their assistance in developing the methodology of this study.

Thanks also go to MSG Lynn Routsong, of the Defense Manpower Data Center, Monterey, California and Helen Davis, of the W. R. Church Computer Center, Naval Postgraduate School, Monterey, California, for their assistance in acquiring, and developing the data set for the study.

I. INTRODUCTION

A. GENERAL

Retaining qualified soldiers in the military after their terms of service are complete continues to be one of the key issues in the all-volunteer Army. Reenlisting good soldiers protects the military's extensive investment in training, and provides the stream of soldiers needed for leadership and supervisory positions. Reenlistments are also a powerful force alignment tool for the Army to balance job skills and grade structure. Although there are many ways for personnel managers to influence reenlistment behavior, the reenlistment cash bonus continues to be the most powerful and responsive tool available.

The United States military has utilized reenlistment bonuses since the early 1960's to improve retention in the services. Since 1974, however, the reenlistment bonuses have been "selective", targeted at specially designated military job skills. To assist military personnel managers in determining which job skills should receive reenlistment bonuses, a large-scale optimization model was developed and refined at the Naval Postgraduate School [Ref. 1: pp. 1-3]. This mathematical model recommends a set of bonuses that attempts to minimize the expected deviation from a desired force structure under the constraint of a given budget. A brief description of this military reenlistment bonus model is in Appendix A.

Use of the military reenlistment bonus model by the U. S. Army is currently limited because of the inadequacy of one of the model inputs, the predicted reenlistment rates. These rates estimate the number of soldiers who will reenlist for each different job skill at each potential bonus level.[1] The military reenlistment bonus model uses these as inputs to determine the most effective method to spend the limited bonus budget.

The purpose of this study is to develop a model to estimate the reenlistment bonus response rates for U. S. Army enlisted personnel for use in the military reenlistment bonus model.

---

[1] It is important to understand that bonuses are a treatment, whose effect on the soldier population is uncertain.

B. BACKGROUND

Reenlistment cash bonuses are executed in the U. S. military through the selective reenlistment bonus (SRB) program. The "selective" bonuses are targeted at specially designated military occupation specialities (MOS) and year-of-service interval (zone) combinations. The U. S. Army currently has over 350 different MOS's. Year-of-service intervals are broken into three zones as follows:

Zone A	2-6 years-of-service
Zone B	6-10 years-of-service
Zone C	10-14 years-of-service [2]

MOS and zone combinations are called cells, and there are over 1000 cells to which the military reenlistment bonus model assigns bonus multipliers. The cash amount of a bonus is computed as follows in Equation 1, where SRB is the cash bonus amount, MBP is the soldier's current monthly base pay, YR is the number of years the soldier reenlists for, and $MULT_{ij}$ is the bonus multiplier for MOS i and zone j.

$$SRB = MBP \times YR \times MULT_{ij} \tag{1}$$

One half of the bonus is paid as a lump sum on the day the soldier reenlists. The remainder is paid in equal yearly installments over the duration of the reenlistment term. Bonus multipliers range between zero and six, and although public law allows them to take on continuous values, the Army restricts them to increments of 0.5. At any given time, 15-25% of the 1000 cells have non-zero bonus multipliers, and the Army's yearly budget for the bonus program is from $50-100 million.

The U. S. Army is currently experimenting by allowing bonus multipliers to vary by rank within an MOS and zone combination. For example, an infantryman in Zone A who achieves the rank of sergeant could receive a higher bonus than soldier of the rank of specialist, a lower rank.[3] The purpose is to encourage more high quality soldiers to reenlist.[4] This experiment causes the bonus multiplier to have three dimensions, ($MULT_{ijk}$) of MOS, zone, and rank. While this study does not address the issue of rank as a dimension of the bonus multiplier, the method outlined here is adaptable to this approach.

---

[2] Soldiers with under two or over fourteen years-of-service are not eligible for reenlistment bonuses. Zone A is extended sightly, to allow soldiers who enlist for two years an opportunity to reenlist prior to the end of their service term.

[3] The rank of sergeant is pay grade E5. The rank of specialist is pay grade E4.

[4] The assumption is that rank is a good measure of soldier quality, an assumption that is used in this study.

Soldiers enlist in the military by signing a contract that obligates them to specific terms of service (usually two to four years). As they near the end of their enlistment term, soldiers have available to them the following options:

REENLIST	A soldier signs a new contract, obligating him or her to a new term of two to six years. Bonuses are for reenlistments of three years or more, and the length of the reenlistment affects the amount of the bonus payment.
REENLIST/MIGRATE	Soldiers also may reenlist, but migrate to a new MOS. Normally this is from an overstrength to an understrength MOS. Usually, migrating soldiers do not receive bonuses.[5]
EXTEND	Extending soldiers defer their reenlistment decision. Extensions are for up to two years, and soldiers do not receive bonuses for extending. Many soldiers extend because they are currently ineligible to reenlist, and they try to become eligible during the extension period. Other soldiers extend to wait for more favorable bonus multipliers. Soldiers also extend to meet schooling, training, deployment, overseas assignment or retirement time remaining in service requirements. Because they are a deferred reenlistment decision, extensions are a major complicating factor to this study. They are addressed in Appendix B.
ETS	End of term of service. A soldier who does not make any of the above decisions is discharged from the service at the end of the contract period.

Soldiers are allowed to reenlist up to eight months prior to the end of their current term of enlistment. Like extensions, this policy also clouds the issue of who is eligible to reenlist at any given time. This issue is also addressed in Appendix B.

The above discussion serves to highlight a few important aspects of the SRB program. For a more detailed overview of the reenlistment system, consult "The Effects of Selective Reenlistment Bonuses on Retention," by Donald J. Cymrot [Ref. 2: pp. 4-9].

C. RESEARCH QUESTIONS

The purpose of this section is to provide the motivation for the specific research areas that will be pursued during this study.

---

[5] Migrating soldiers can expect faster promotion rates in their new shortage MOS.

1. MOS Grouping

This study is sponsored by the U. S. Total Army Personnel Command, Alexandria, Virginia. Their task is to develop a model to estimate reenlistment response rates for use in the military reenlistment bonus model. A brief review of the input form required by the bonus optimization model motivates the approach of the study. Figure 1 shows a graphical example of the input requirement for the military reenlistment bonus model.

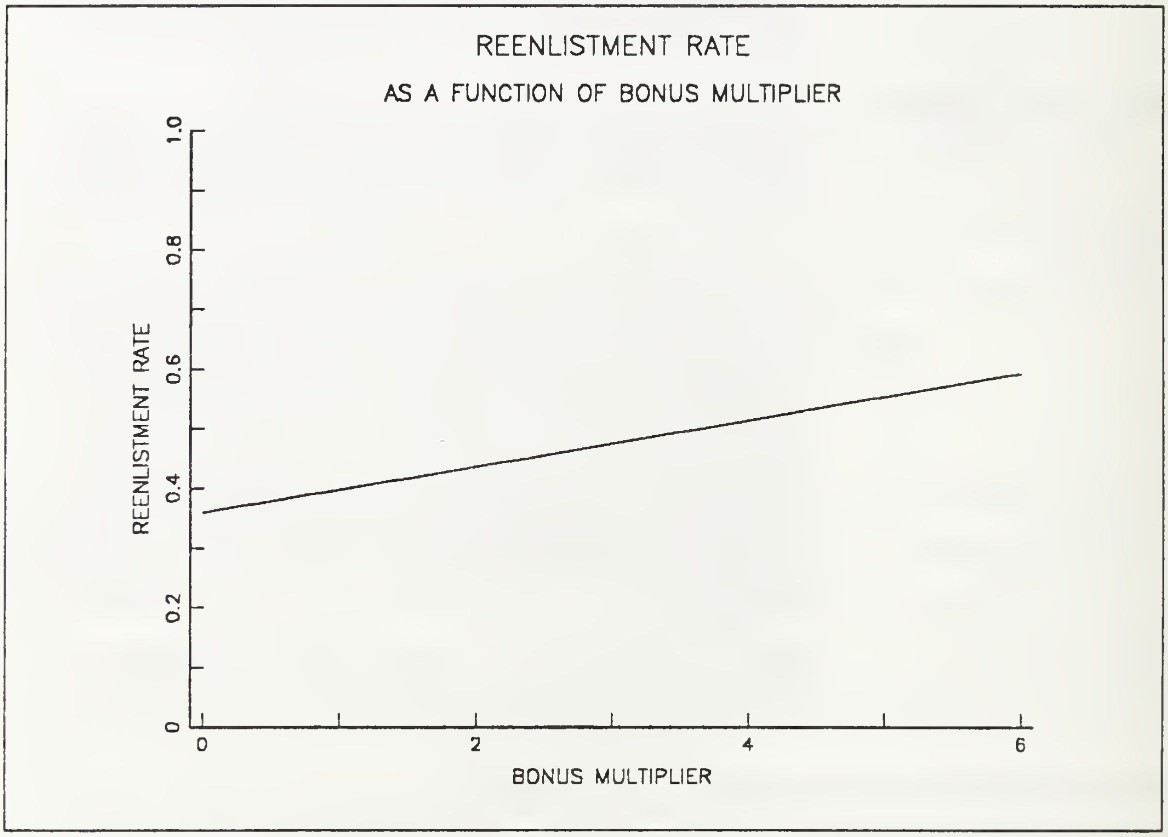

Figure 1. Sample of Input Required for Bonus Model (Hypothetical)

The military reenlistment bonus model requires as input a function that takes a specified bonus level and outputs the expected reenlistment rate, by MOS.[6]

A point to note is that the above example is MOS and zone specific. The bonus optimization model requires over 1000 such functions (one for each cell). However, the computer resources are not available to execute the 1000 different regression models necessary to develop the 1000 different response functions. The goal of this study is to develop a methodology to reduce the number of regression models, by some appropriate grouping technique.

---

[6] The actual function is input into the military reenlistment bonus model as a point estimate for each of the various bonus levels.

A brief review of past attempts at grouping of MOS's gives some perspective to this research question. The first attempts at grouping combined all MOS's together. They estimated one set of reenlistment response rates for all MOS's. One study taking this approach is Enns [Ref. 3: pp. 1-3]. The problem with this approach is that there is evidence of the varying effects of reenlistment bonuses among MOS's. The strongest evidence of this is found in research by Lakhani and Gilroy [Ref. 4: p. 253].

The next attempt was to estimate a separate reenlistment response for each different MOS. In addition to the problem noted above (the requirement for 1000 different regression equations), there are a number of additional problems with this approach. The first problem is that since bonuses are allocated by MOS, it follows that all soldiers within the same MOS (and zone) receive the same bonus [Ref. 5: p. vi]. This limits the number of observations at different bonus levels available for use in the regression. To further complicate this problem, only 15-25% of the over 1000 cells have non-zero bonus multipliers at any given time. Large numbers of cells never have a bonus, or have such a limited bonus history that estimation by regression techniques is meaningless.

A second problem with estimating a separate reenlistment response rate for each MOS is that bonuses within a speciality often do not change from year to year. This is caused by the fact that bonuses are often given to critical MOS's, and these MOS's remain critical over time. One study by Hosek and Peterson [Ref. 6: pp. 19-22] estimates the correlation of bonus levels in adjoining time periods to be 0.8 for specialities receiving a bonus. This correlation causes the regression model to behave poorly.

A third problem is that this technique assumes the MOS is a homogeneous grouping of soldiers with similar reenlistment probabilities. However in his research, Kohler questions this assumption and shows that MOS's are not homogeneous groupings [Ref. 5: p. 4].

To correct for the deficiencies with estimating reenlistment response rates, most researchers have grouped MOS's. The advantage to this approach is that by grouping MOS's with varying bonus levels together, the regression estimates become more meaningful. Two basic approaches are used. The first approach is to group MOS's into career management fields (CMF's). The Army currently has 32 CMF's. Studies using this technique include a study of Army reenlistment and extension decisions by Lakhani and Gilroy [Ref. 4: p. 232]. The problem with this approach is that the CMF's are administrative groupings, and CMF's often group occupations with little in common [Ref. 5: p. 4].

The second approach is to assign MOS's into groups with similar job characteristics. These characteristics tend to key on how technical is the job, what is the skills potential combat exposure, or what are the skills civilian opportunities. Presented below is a listing of groupings in the Concepts Analysis Agency (CAA) bonus study [Ref. 7: p. 4-21].

- Direct combat
- Combat operations
- Communications/electronic operations
- Communications electronic maintenance
- Mechanical maintenance
- Supply services transportation
- Medical
- Administration
- Engineer Construction
- Intelligence

Groupings such as these make intuitive sense. However, analysis supporting use of these groupings is lacking. The key point is the goal of grouping is not only to reduce the number of regressions to be performed, but also to form groups with similar reenlistment behavior. Therefore, to improve the quality of the estimates of reenlistment response rates, this study develops techniques to identify groupings of soldiers with similar reenlistment probabilities.

2. Variables to be Considered

The study of the effects of reenlistment bonuses is not a trivial problem. It is difficult to determine why soldiers decide to stay or leave the service. There are many factors which impact a soldier's reenlistment decision, as diverse as what the job opportunities in his hometown are, to whether he is well adjusted within his organization, to what the congressional action is on pay raises for the next year. The reenlistment decision is based not only on the bonus offered, but upon many other factors, both quantifiable and unquantifiable. The impact of these other factors is seen in Figure 2, which is a scatterplot of quarterly reenlistment rates for ten different Zone A MOS's over four years, as a function of the bonus level. Although there is a general increasing trend in the reenlistment rate, many other factors are working to produce the observed variance. Without the explanatory effect of other variables, it is difficult to determine the true effects of reenlistment bonus.

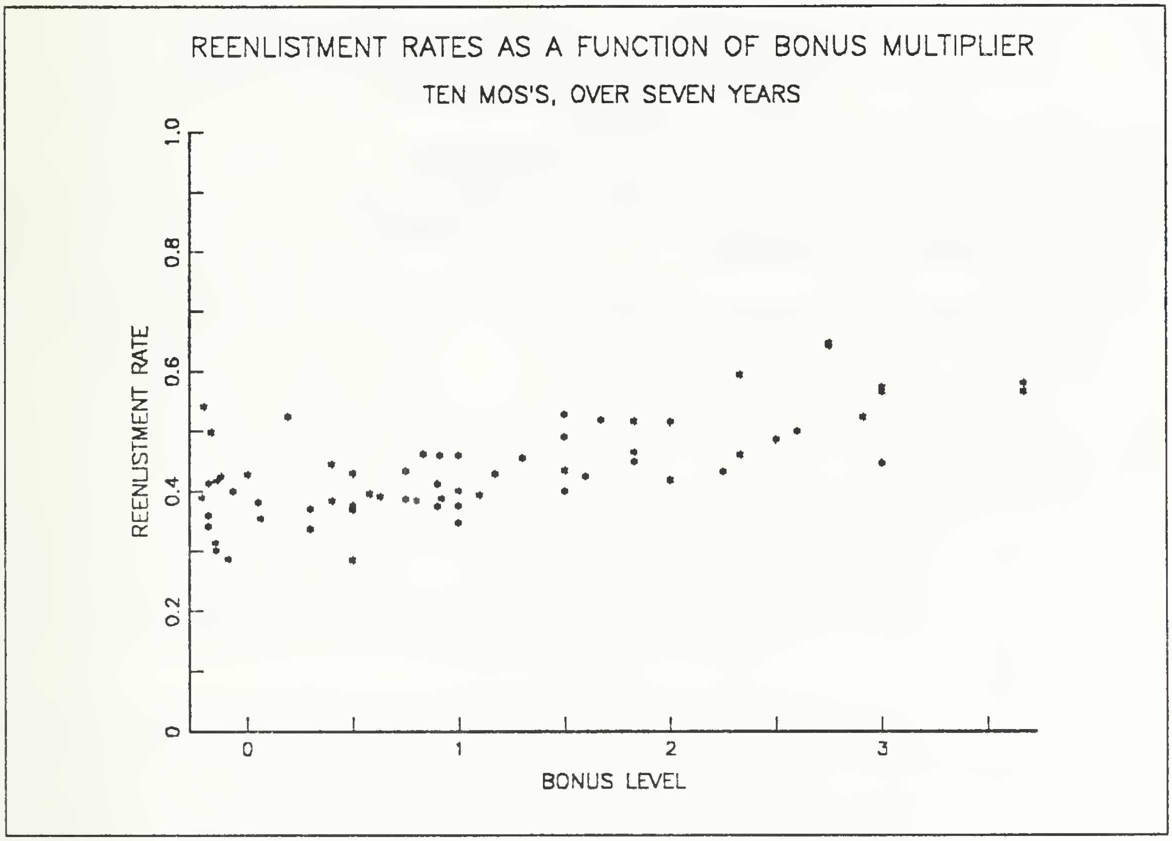

Figure 2. Yearly Reenlistment Rates for Ten MOS's Over Seven Years

Many researchers fail to examine the full range of potential, quantifiable explanatory variables available. For example, the 1982 CAA study uses only three explanatory variables: the bonus level, unemployment, and the inflation rate [Ref. 7: p. 4-10]. Only two studies, a study by Chow and Polich [Ref. 8: pp. 29-31] and a study by Hiller [Ref. 9: pp. 20-31] examine a full range of variables.

This study examines a full range of potential, quantifiable explanatory variables. First, a theoretical framework of the reenlistment decision making process is developed. This framework guides the selection of variables and the gathering of data. Exploratory data analysis techniques are used to determine which of the variables are most appropriate for inclusion in the regression equations. Cross-validation is used to lend credibility to this analysis.

Special attention is paid to the effects of variables that the Army manipulates to influence retention. Variables the Army manipulates in this manner are called force alignment variables.

3. Summary of Research Questions

In summary, the following are the primary research questions of this study.

- Which variables to include in the models?
- How do force alignment variables impact reenlistment?
- How to group soldiers to reduce the number of regression models required, and ensure homogeneous groupings?
- How to address MOS migration and extensions, along with reenlistment eligibility requirements without complicating the model?
- What confidence to place in the estimates?

D. SCOPE OF THESIS

Due to the stated purpose of this study, research is limited to active duty U. S. Army enlisted soldiers, with between 2 and 14 years-of-service. Within this framework, the emphasis is placed on Zone A reenlistments,[7] as the large majority of the bonus recipients are in Zone A.

Because of the extensive research conducted in this area, an attempt is made to draw on previous studies to put together a comprehensive study of estimating reenlistment behavior for the U. S. Army. However, because of the requirement to estimate coefficients for all MOS's, individual MOS differences which warrant special attention are for the most part ignored.

One final note. This study does not address the issue of quality of the reenlisting soldier. Because the military reenlistment bonus model does not distinguish between soldiers, all soldiers qualified to reenlist are assumed to be of equal quality.[8]

E. ORGANIZATION OF THESIS

Chapter II is a review of the literature relevant to the estimation of reenlistment response rates.

---

[7] Zone A extends from 2-6 years-of-service (YOS), Zone B from 6-10 YOS and Zone C from 10-14 YOS.

[8] The experiment outlined in the introduction, (page 2) which treats rank as a separate dimension, attempts to address the quality issue. However within the new cell (dimensioned by MOS, zone and rank), all soldiers are considered of equal quality and the same assumption is made here.

Chapter III develops a theoretical framework for the reenlistment process, and the data base is structured using this framework.

Chapter IV describes the solution technique.

Chapter V shows, in detail, the solution of the Zone A problem. Chapter V also discusses the validation of the Zone A model and the precision of the model. Chapter VI gives the conclusions and recommendation for further study.

The appendices contain various details of interest to the reader, including background on the military reenlistment bonus model, details on how the study details with factors such as MOS migration and extensions and issues such as variable selection, data set cleaning, regression models and statistical tests.

F. STATISTICAL PACKAGES

The statistical package used in this study is SAS, by the SAS Institute. Graphics was done using a pre-release version of GRAPHSTAT by IBM.

II. REVIEW OF THE LITERATURE

A. GENERAL

The purpose of this chapter is to review the literature on the estimation of reenlistment rates, with the purpose of providing motivation for the techniques of this study. The issue of reenlistment bonuses is well studied; this review addresses only a portion of the work done.

B. ARMY STUDIES

The 1982 Concepts Analysis Agency (CAA) study addresses both a method for optimizing bonus payments, and estimates of reenlistment bonus response rates [Ref. 7: p. 4-16]. The study calls these rates SRB effectiveness coefficients, and the coefficients they estimated in 1982 are still in use today by the Force Alignment Branch of the U. S. Total Army Personnel Command.

The CAA study uses 1976-1981 data and variables to measure the bonus level, the unemployment rate, and the inflation rate. Over 320 MOS's are grouped into ten skill groups,[9] and linear regression models are used to estimate the SRB effectiveness coefficients.[10] The study does not estimate reenlistment rates, instead it recommends use of the current reenlistment rate as the forecast reenlistment rate.

A second study of Army bonus response rates, by Higham [Ref. 10: pp. 9-13], uses linear regression and variables that measure the bonus level, year, calender quarter, unemployment rate and inflation rate to estimate reenlistment rates. The study estimates reenlistment rates for twenty-four MOS's with good bonus histories, and then describes techniques to extrapolate the results to the remaining 300 MOS's.

Both of these studies use linear regressions; Appendix I explains why logistic regression is preferred over linear regression in studies such as these. Both studies also examine a limited number of explanatory variables. One of the goals of this study is to examine a large number of variables for inclusion in the model. Neither study presents cross-validation results for their models. This study uses cross-validation to ensure model fit.

---

[9] These skill groups are listed on page 6

[10] The SRB effectiveness coefficients are the percentage increase in the reenlistment rate due to a one step increase in the bonus multiplier.

Another study of reenlistment propensities has been done by economists of the Army Research Institute for the Behavioral and Social Sciences [Ref. 4: pp. 229-232]. The study uses bonus levels, a civilian military wage index, the unemployment rate, the soldier's AFQT score[11], race, family size and groups soldiers by career management field. This study is interesting in two respects. First, it examines three choices in the reenlistment decision making process, and therefore applies multinomial logistic regression. The three choices are to reenlist, to extend, or to leave the service. Researchers are split over whether to treat the extension decision as a separate choice, or to treat it as a deferred reenlistment decision. Our study chooses to treat extensions as a deferred reenlistment decision. Appendix B gives further explanation and justification.

A second interesting aspect of the study is the grouping of MOS's into career management fields.[12] Many MOS's do not have adequate enough bonus histories for regression models. Therefore, most studies group MOS's, either into career management fields or into groupings with similar job characteristics. A goal of our study is to examine an alternative grouping technique, in which soldiers are grouped according to their reenlistment probabilities, regardless of which MOS's they are in.

A final Army study discussed here is by two economists at the United States Military Academy [Ref. 11: pp. 211-212]. This study points to the examination of demographic variables, such as race, sex, and family size as the method to form homogeneous groupings of soldiers with similar reenlistment probabilities. This method is followed in Chapter V of this study.

C. ACOL STUDIES

The Navy has done extensive research into the prediction of reenlistment response rates. The annualized cost of leaving model (ACOL) represents the current state of the art of its research [Ref. 12: pp. 2-5]. ACOL models the reenlistment decision making process by examining the present value of the soldier's military pay potential and his or her civilian pay potential. It also examines the soldier's "taste for military service". The model has a great deal of potential; however, it does carry some difficult to validate assumptions, such as the time horizon over which a soldier makes a decision, his or her discount rate, what their civilian earnings potential is, and whether the soldier's perceptions of his or her earning potential is close to realistic.

---

[11] AFQT is the Armed Forces Qualification Test

[12] Career management fields are an administrative grouping of MOS's used by personnel managers to administer personnel programs.

One study that uses this ACOL methodology is a Marine Corps study by Cymrot [Ref. 2: pp. 24-25]. Cymrot groups marines into twenty-two skill families, and uses the one year difference between the military pay and civilian pay potential, along with variables to measure the bonus level, the unemployment rate, and the current rank of the soldier.

The ACOL model holds a great deal of potential for predicting reenlistment rates. However for reasons of scope and data availability, it is not fully incorporated into this study. Instead, variables that measure the first year difference between civilian and military wages are included in this study, in a manner similar to the Cymrot study approach.

This brief review of the literature services to further motivate the research questions introduced in Chapter I. Additional review of the literature appears in Chapter III.

III. DATA BASE

A. GENERAL

One of the shortcomings of many previous reenlistment studies is that they fail to consider a broad range of variables which may explain reenlistment behavior. For example, the 1982 Concepts Analysis Agency study examines only three explanatory variables: the bonus level, the inflation rate, and the unemployment rate [Ref. 7: p. 4-10]. One of the goals of this study is to examine a full range of potential, quantifiable explanatory variables.

The purpose of this chapter is to describe the selection of variables and the development of the data base. A conceptual framework is developed to give focus and direction to the data gathering effort. At this point, it is not important to assess the potential significance of any particular variable, or to establish relationships between them; instead it is sufficient to create a list of promising variables. In Chapter V, exploratory data analysis techniques determine which variables to include in the regression equations. Seven variables are included in the regression model.

This chapter focuses primarily on the conceptual framework for the Zone A reenlistment decision.

1. Source of Data

Data for this project comes primarily from the Defense Manpower Data Center (DMDC), in Monterey, California. The mission of this organization is to archive manpower data from all services for use in studies such as this. The Army gain loss file is the primary source of data for the project. Other data includes economic variables from sources such as the Bureau of Labor Statistics.

The data available from DMDC are records of soldiers actually making reenlistment decisions. Individual-level records are chosen for the analysis rather than group-level data because the later provides only limited insight into which variables influence soldier retention. To study the determinants of reenlistment behavior, data on individuals themselves are most appropriate [Ref. 13: p. 3]. However, the analysis of individual-level data is not without its costs in computing time and data storage requirements.

2. Response Variable

The response variable for the study is binomial: either the soldier chooses to reenlist in his or her MOS or not. Some studies model the reenlistment decision-making process as a multinomial choice of reenlistment, extension, or leave the service. Appendix B addresses the issue of why a binomial response variable is chosen over a multinomial response variable.

3. Explanatory Variables

This study includes a variable in the data base if it is quantifiable and if there is some indication (hypothesized or in previous literature) that this factor explains the reenlistment decision-making process.[13] The ideal variable is one that is also predictable in the future [Ref. 14: p. 20]. In those cases where a primary variable is not quantifiable, the study develops surrogate variables. For example, it is difficult to quantify the success of a soldier. This study uses the rank the soldier achieves and the speed with which he achieves it as surrogates for military success.

4. Survey Data

Survey data is not included in the data set. Unfortunately, this eliminates the only way to measure a considerable number of reenlistment factors, especially those concerning soldier attitudes towards their jobs, and living conditions. However the problems with survey data are twofold. First, it is impossible to match survey data with the individual records. Second, although some past surveys are available, the survey effort falls considerably short of the scope of the individual data gathering effort. Survey data, and the studies that analyze it, assist in providing the insight necessary to choose variables for this study. However, survey data is not available to measure those variables.

5. Time Period Covered

The data base covers the period from the fourth quarter, FY80 thru the first quarter, FY89, 34 quarters of data in all. Data obtained before 1980 are not included for practical reasons. Prior to that date, DMDC stored data in the gain loss file in a different format than is used at present. Conversion of that data is an expensive, time consuming process, which is not justified for this project.[14]

---

[13] If a variable explains the reenlistment decision-making process it means that it reduces the uncertainty of prediction of reenlistment rates.

[14] One advantage to including more data (prior to 1980) in the study is to improve the range of values of the explanatory variables. However, analysis shows that all variables have a good range of values, and only modest improvement is achievable by including values from 1974-1979. A

6. Size of Data Set

The data set contains the records of over 500,000 Zone A soldiers making their reenlistment decisions. The study breaks the data into two groups, one group of data for analysis and development of the regression models, and the second group of data for validation. Numerous previous studies have neglected the validation process; the latter step is a requirement for lending credibility to any predictive model.

B. CONCEPTUAL FRAMEWORK

We hypothesize that the reenlistment decision-making process of a soldier considering reenlisting for the first time depends on the following four factors.

- The soldier's initial motivation for military service.
- The soldier's success in the military and satisfaction with military life.
- The soldier's evaluation of the potential for success outside the military.
- The influence of Army reenlistment policies on the soldier's initial decision to stay or leave.

First some comments on the specifics of this framework.

1. Initial Motivation for Military Service

Previous research supports the hypothesis that initial enlistment motivation influences a soldier's first term reenlistment behavior.[15] For example, an Air Force study of first-term reenlistment intentions of avionics technicians lists career intentions at the time of enlistment as the most important factor contributing to the technician's reenlistment plans [Ref. 15: p. vii]. Of course the difficulty is measuring enlistment motivation. The most direct way is to survey soldiers; however, historical survey data is not available. Instead, this study uses the following variables to gain insight into enlistment motivation.

- Army College Fund Program Participation (ACF)
- Enlistment Bonus
- Enlistment Term
- Enlistment Program/Training Program
- Age at Enlistment

---

second reason not to include data prior to 1980 is relationships between explanatory variables and dependent variable may change over time; emphasis is best placed on the more recent history.

[15] The terms Zone A and first term are interchangeable in this study. Both refer to soldiers making their first reenlistment decision, usually after two to four years of service.

- Age at Separation
- Education at Enlistment
- Dependent Status at Enlistment
- Prior Service
- Reserve Time
- Youth Program
- Hometown
- Unemployment Rate at Time of Enlistment

The study uses these variables to determine whether a soldier is job, training or education-motivated. While these variables do not directly measure a soldiers enlistment motivation, they give insight into it, which in turn helps predict the soldiers reenlistment propensity.

Appendix C gives a detailed discussion of each of these variables.

2. Success in the Service and Satisfaction with Military Life

The soldier's motivation for entering the service determines his or her initial reenlistment propensity. However, the success the soldier achieves in the first term, and his or her satisfaction with military life, profoundly effects this initial reenlistment propensity. As before, there are problems with directly measuring these factors. For example the military uses items such as enlisted evaluation reports, skill qualification tests, awards, and promotions rates to measure a soldiers success. Of these, only promotion rate information is available for use in this study. However, at least numerous studies support using promotion rates as a measure of success in the military. In one study by Ward [Ref. 16: p. v] promotion speed relative to that of peers is the only indicator of a high level of achievement. Two studies go further and try to predict promotion rates using intelligence and educational scores. Although the results of these studies are not consistent nor particularly strong, this study includes intelligence and educational variables [Ref. 16: pp. 1-3] [Ref. 17: p. 14].

Measuring a soldier's satisfaction with military life is also difficult. However numerous studies find that quality of life issues appear to have little effect on the first term reenlistment decision, although the impact of these factors increase dramatically in importance thereafter. For example, one study uses survey data to show that although military families do not like separations, they do not leave the service because of them [Ref. 18: p. 27]. Supporting this is a study which finds the effects of factors such as family separations are not significant in the first term reenlistment model [Ref. 8: p. 25].

Two studies by the Navy Personnel Research and Development Center find that quality of life issues are not statistically significant predictors of first term reenlistment intent [Ref. 18: p. vii] [Ref. 19: p. vi]. One quality-of-life issue that has some significance is first term duty location. One researcher finds that soldiers stationed overseas during their first-term have reenlistment rates higher then those stationed in the continental United States [Ref. 8: p. 23].

As a result of the above arguments, this study includes the following variables.

- Character-of-Service
- Promotion Rates
- AFQT Score
- Mental Test Category
- GT Score
- Education Level at Reenlistment
- Change in Education
- Years-of-Service
- Current Rank
- Duty Location
- Dependent Status at Reenlistment
- Change in Dependent Status

Appendix D discusses each of these variable in more depth and provides further motivation for including them in the analysis.

3. Evaluation of Potential in the Civilian Sector

We are developing a conceptual framework to explain the reenlistment decision-making process of soldiers. The framework starts by looking at the soldier's initial enlistment motivation. This motivation (whether it is job, training or education) gives the soldier an initial bias towards staying or leaving the service. The soldier's initial bias is changed based on the success the soldier achieves in the first enlistment term and his or her adjustment to military life. Many soldiers decide during the first term that the Army is not for them, and they leave the service. However, we hypothesize that many soldiers decide whether to stay or leave the service after making a comparison of their military and civilian potential. The purpose of this section is to discuss the variables associated with this comparison.

An issue is whether soldiers can make meaningful evaluations of their potential in the civilian sector. This study assumes they can. Secondary issues are: how can the study measure the soldier's opportunities, and does the study's evaluation of a soldier's potential match the soldier's evaluation of his or her potential?

There are a number of ways to measure the civilian opportunities available to a soldier. One way is to look at the job category the soldier is in, and employment growth of comparable civilian jobs. Another is to look at the civilian military wage index. These efforts are hampered due to incompatibility of numerous Army skills with comparable civilian skills. Additionally, national economic indicators such as gross national product (GNP), consumer price index (CPI), and the unemployment rate to are used to assess the civilian opportunities available to the soldier.

Finally, the study uses demographic variables as surrogates for the civilian versus military evaluation a soldier makes. Researchers note that women and black soldiers reenlist at higher rates than white male soldiers. The researchers hypothesize that this is due to women and blacks seeing insufficient job opportunities in the civilian sector, as compared to military career options. Additionally, researchers hypothesize that women and blacks see enhanced promotion opportunity in the military as compared to the civilian sector. [Ref. 14: p. 29]

The study therefore uses the following variables to explain the soldier's evaluation of potential in the civilian sector:

- Race
- Ethnic Group
- Sex
- Job Type
- Unemployment Rate
- Civilian Military Wage Index
- Consumer Price Index
- Gross National Product
- Percentage Growth Civilian Jobs

Appendix E describes each of the above variables in more depth and provides further motivation for including them in the study.

4. Reenlistment Policy Variables

After soldiers compare opportunities in the civilian sector to those in the military, they make an initial reenlistment decision. However, the impact of Army reenlistment policies can change this decision. For example, a soldier who initially decides not to reenlist may change his mind in response to the offer of a reenlistment cash bonus. A soldier who initially wants to reenlist may change her mind because she is unable to get the reenlistment option of the training or duty station she desires. Additionally, changes in reenlistment eligibility may make the soldier ineligible to reenlist. The above are examples of the affects of reenlistment policy variables.

The Army is not able to directly manipulate all variables listed in this section. For example, military pay and the retirement programs are policies that the Army can only recommend to Congress. However, all the variables in this section are policy variables at some level in the government.

The study includes the following policy variable:

- Retirement System
- Number of Years to Military Retirement
- Real Military Compensation (RMC)
- RMC Adjusted by Inflation
- Bonus Payment
- Type of Bonus Payment
- Job Skill Migration
- Promotion Rate Forecast
- Reenlistment Eligibility Criteria
- Reenlistment System

Appendix F discusses each of these variables in more depth and the motivation for including each of them in the analysis.

C. SIGNIFICANCE OF UNQUANTIFIABLE VARIABLES

Despite including over forty variables in this study, there are still numerous unquantifiable factors which may explain the reenlistment decision-making process. Those related to satisfaction with military life appear to have little effect on the Zone A decision. However this study also excludes job satisfaction variables, such as autonomy, physical work environment, skill utilization, team effort, and relationships with peers, subordinates and supervisors. This is unfortunate, because studies show job satisfaction is extremely important for the first term reenlistment model[16] [Ref. 20: p. iii]. Job satisfaction variables are excluded because they are not measurable, except by survey, and survey data is not available in sufficient detail to match the study's data set. Additionally, job satisfaction variables are difficult to predict (forecast) and therefore do not fit well in the reenlistment model.

What is the significance of omitting variables such as job satisfaction? More unexplained variance may appear in the regression models, which leads to less precision and confidence in the reenlistment response rates. We discuss these issues in more depth later.

D. CLEANING THE DATA SET

Initial study indicates that the data set has a considerable amount of inaccurate data. For example, Figure 3 shows the variable **TERM OF ENLISTMENT**. For this variable, 6.1% of the entries are for zero or one years, or for more than four years, which are invalid terms of enlistment.[17] Analysis shows that invalid data rates range from 0-15% for most variables; however, seven of the variables have error rates of 15-25%.[18]

Clearly there is a need to investigate the source of the data errors, and determine the potential impact on the analysis. This investigation revealed that every entry for FY81 is in error for the seven variables with error rates of 15-25%. Discussions with DMDC determined that the data file used in this study was a merging of two other data files, and in the case of FY81, this merging was incorrectly performed. While DMDC is correcting the problem for future use, the corrections were not available for use in this study. Therefore, FY81 data were excluded from further analysis.

DMDC referred us to the U. S. Total Army Personnel Command for an explanation of the error rate of up to 15% on the remaining variables. The information systems managers acknowledged that they had difficulty obtaining accurate data from Army organizations, and although they said efforts are underway to improve the quality of the data, they offered few suggestions of how we could improve our data set.

Rather than discard all records with invalid data, an attempt was made to clean the data set by cross referencing other data. An example is the variable **TERM OF**

---

[16] However, job satisfaction decreases in importance in the second term.

[17] Inaccurate data are determined by consulting the appropriate Army Regulation for the acceptable ranges of entries.

[18] There is no missing data in the data set.

ENLISTMENT Figure 3 shows the errors in this variable for a random sample of 75,778 records.

TERM OF ENLISTMENT

			CUMULATIVE	
TERM	FREQUENCY	PERCENT	FREQUENCY	PERCENT
0	4405	5.8	4405	5.8
1	36	0.0	4441	5.9
2	5760	7.6	10201	13.5
3	42853	56.6	53054	70.0
4	22577	29.8	75631	99.8
≥ 5	147	0.2	75778	100.0

Figure 3. Frequency Counts for the Variable Term of Enlistment, Uncleaned

TERM OF ENLISTMENT values of zero and one year are not valid, nor are values of greater than four years. The study corrects for this by examining enlistment dates and reenlistment dates and inferring from this the enlistment term. Following cleaning, the variable **TERM OF ENLISTMENT** has the distribution of Figure 4.

TERM OF ENLISTMENT

			CUMULATIVE	
TERM	FREQUENCY	PERCENT	FREQUENCY	PERCENT
2	6291	8.3	6291	8.3
3	44784	59.1	51075	67.4
4	24703	32.6	75778	100.0

Figure 4. Frequency Counts for the Variable Term of Enlistment, Cleaned

Using procedures such as described above, much of the invalid data was corrected. Appendix G lists the amount remaining by variable. Error rates range from 0-7.8%, with numerous variables having less than 1% invalid data. As a part of the cleaning process,all remaining invalid data were recoded as missing data.

The question is whether the amount of missing data listed in Appendix G are acceptable, or if additional cleaning is necessary. The SAS statistical procedures of this study exclude observations with missing values from further analysis [Ref. 21: p. 550]. Therefore, missing values are of concern if they constitute a high percentage of the observations in the multidimensional analysis, or if the missing values are not randomly distributed throughout the observations.[19] However, our analysis shows that the amount of remaining missing data is reasonable, and that the missing data does not change the results of our analysis. Appendix G show the results of the statistical procedures that show these results. Therefore, no further cleaning of the data set is done. Continuous variables are cleaned in a similar manner.

---

[19] An example of non-randomly distributed missing values is the seven incorrectly coded variables of 1981, discussed above.

IV. METHODOLOGY

A. GENERAL

The purpose of this chapter is to motivate the new methodology for predicting reenlistment rates.

B. MOTIVATION FOR THE METHODOLOGY

1. Problems With Current Solution

The purpose of this study is to predict reenlistment rates for each of the Army's 350 military occupation specialities (MOS). However, it is impractical to do a separate regression on each of the different MOS's for a number of reasons. These reasons were discussed in some detail in Chapter I, and are reviewed here.

- Many of the 350 MOS's (60-70%) have never (or infrequently) been assigned a reenlistment bonus. Estimates of regression coefficients for those MOS's produce misleading results, because of the inadequate range of bonus values.
- All soldiers in an MOS receive the same bonus level at the same time, and therefore it is difficult to separate the effects of the bonus level from other explanatory variables.
- Bonus levels have a very high correlation from year to year within an MOS, which degrades the accuracy of the regression results.
- There is evidence that MOS's do not represent homogenous groups of soldiers with similar probabilities of reenlisting. Therefore, considerable variance is added to the problem before the regression is conducted.

Numerous previous studies have addressed these problems by grouping MOS's together, usually forming 10-20 groups of 10-50 MOS's. Grouping in this manner is usually done by combining MOS's that have similar job characteristics. The Concepts Analysis Agency study uses this approach [Ref. 7: p. 4-21].

Forming groupings of MOS's in this manner solves the first three of the four problems listed above. There are, however, two criticisms of this technique of grouping MOS's. First, the groupings are formed on an intuitive basis, and no attempt is made to quantitatively determine if the grouping is sensible. Second, the fourth problem listed above (MOS's are not a homogeneous grouping of soldiers with similar probabilities of reenlisting) is not solved. Clearly, if an MOS is not a grouping of soldiers with a similar probabilities of reenlisting, then neither is a grouping of MOS's.

A major theme of this thesis is analysis of a new technique of grouping soldiers. The methodology looks for groupings of soldiers with similar probabilities of reenlisting, independent of their military occupation specialities. Since the groups contain soldiers of differing MOS's, they have robust bonus histories, and less correlation from year to year. Potentially, this grouping technique solves all four of the problems listed above.

To more fully explain and motivate this solution, the assertion that an MOS is not a collection of soldiers with similar probabilities of reenlisting is now examined.

2. Non-homogenous MOS

Previous research supports the assertion that an MOS is not a homogenous grouping of soldiers with similar probabilities of reenlisting [Ref. 5: p. 4]. This section provides examples to illustrate the point.

First the fact that an MOS has subgroups of soldiers with widely varying reenlistment probabilities is demonstrated. As an example, Infantrymen (MOS 11B) have a 34% reenlistment rate over the past six years. However, when the MOS is partitioned into two categories by **DEPENDENT STATUS** (one category is single soldiers without dependents, and the second category is married and single soldiers with dependents)[20] these two categories display widely varying reenlistment rates of up to 20%. Figure 5 shows the example for Infantrymen (MOS 11B).

This result is not unique. Figure 6 shows three other MOS's which also display the same characteristic. Additionally, Figure 6 shows that all MOS's taken together also display about a 20% difference between the reenlistment rates for soldiers with and without dependents. Although the actual rates differ some by MOS (there are many different factors interacting in this simple example) the general trend holds.

There are other variables that have similar characteristics. For example, Figure 7 shows Infantrymen (MOS 11B) partitioned into categories by **RACE**.

---

[20] Dependents may be children, elderly parents or any other legal dependent

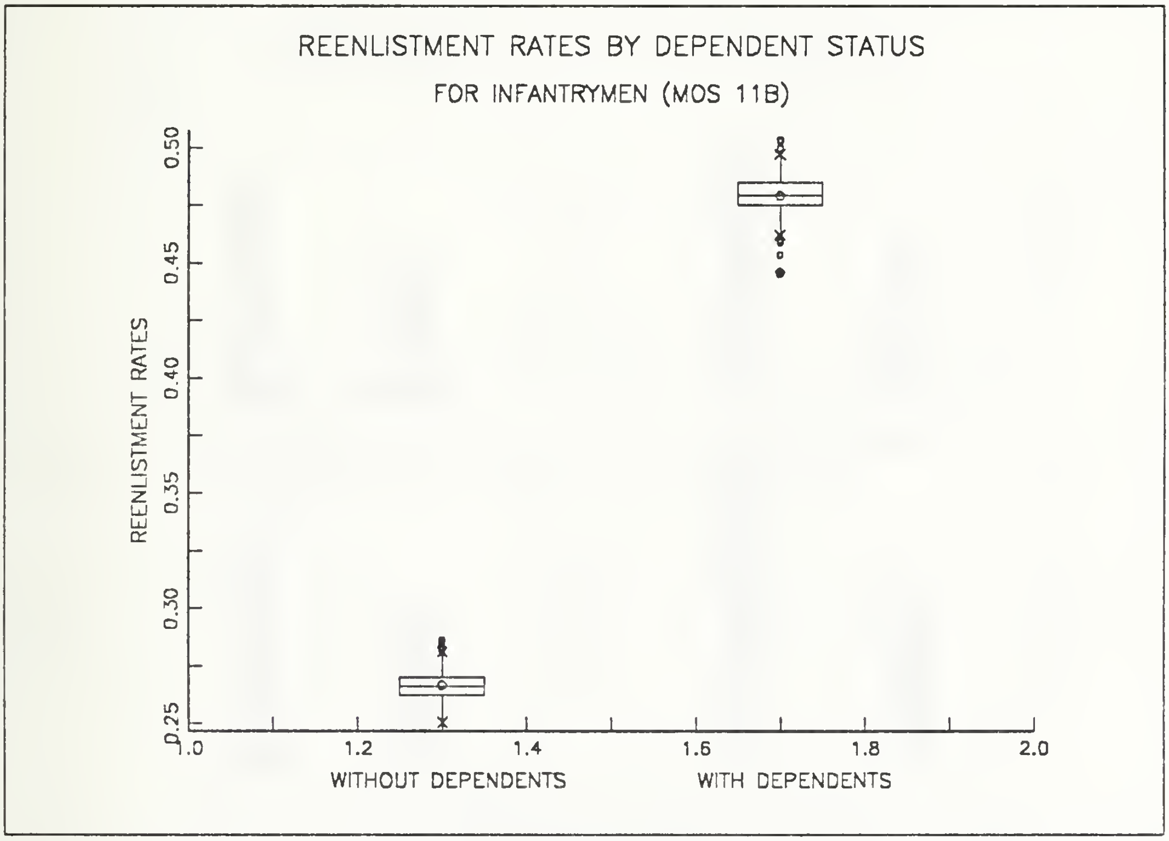

Figure 5. Reenlistment Rates for MOS 11B, Zone A by Dependent Status

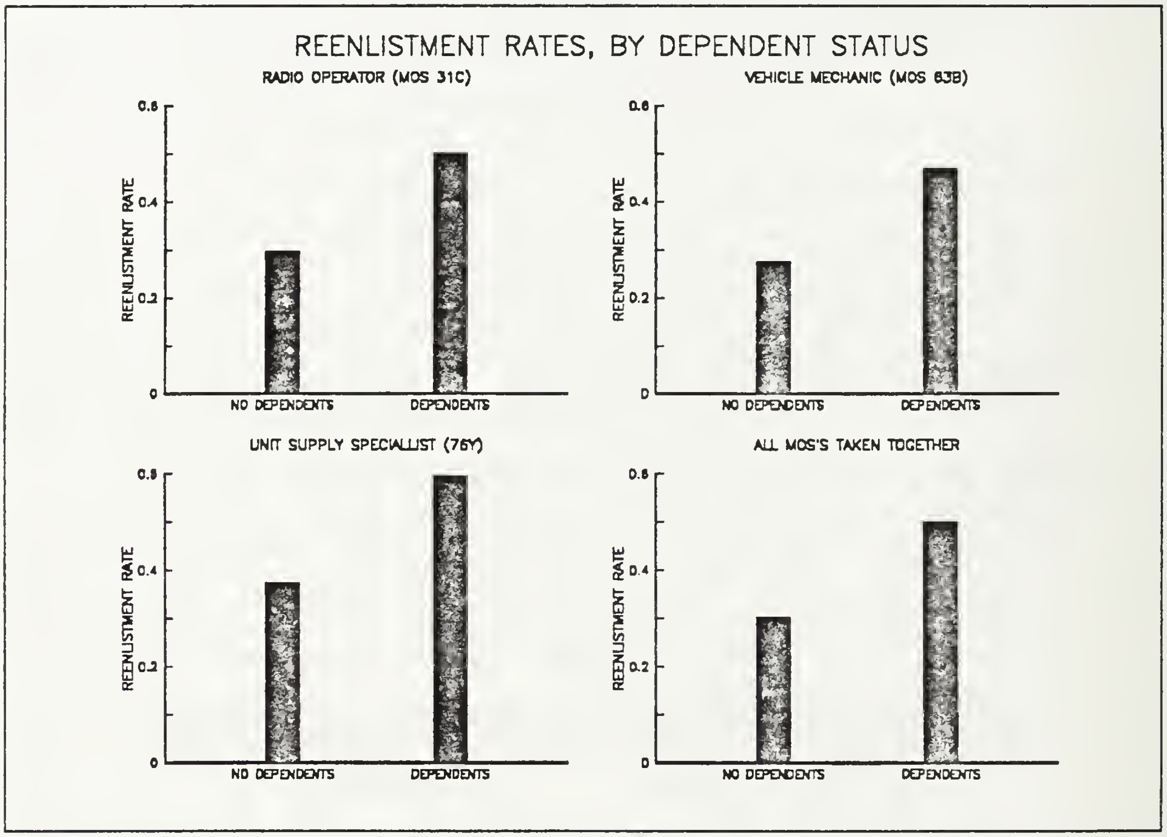

Figure 6. Reenlistment Rates for Differing MOS's by Dependent Status

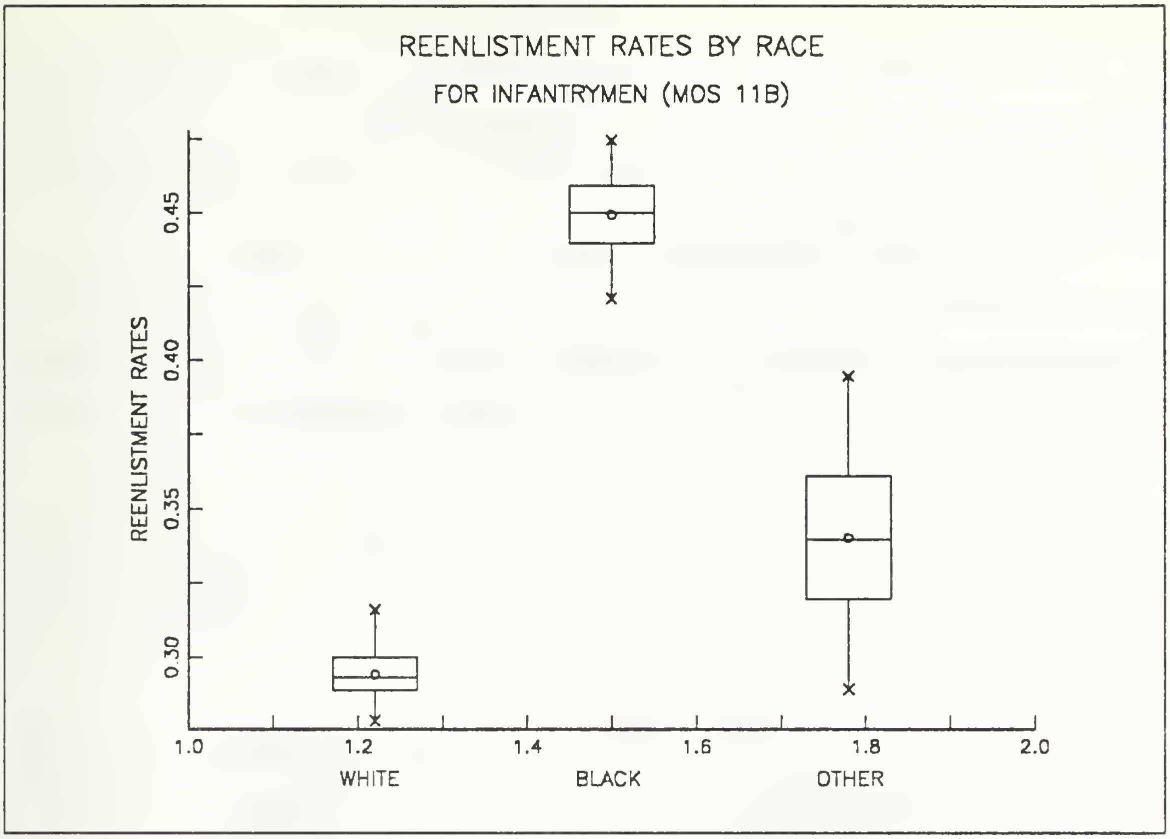

Figure 7. Reenlistment Rates for MOS 11B, Zone A by Race

Clearly, the different racial groups have differing reenlistment rates, by up to 15%. There are many other examples, some of which are summarized in Table 1. Percentages are for all MOS's taken together, and do not necessarily include all categories.

Table 1. REENLISTMENT RATES BY CATEGORY, FOUR VARIABLES

Term of Enlistment	2 Years	19%
	> 2 Years	40%
Sex	Male	37%
	Female	46%
Region of Country	Northeast	27%
	South	49%
Paygrade	E4	38%
	E5	57%

From this simple example it is possible to see that an MOS is not a homogeneous grouping of soldiers with respect to reenlistment propensity. There are categories of the MOS that display widely differing probabilities of reenlisting. These results are seen in most MOS's analyzed.

Once we establish that the MOS is not a homogeneous grouping of soldiers with similar reenlistment rates, we also want to show that different MOS's are comprised of varying percentages of soldiers from the different categories. To illustrate this, a simple example using Infantrymen (MOS 11B), Unit Supply Specialist (MOS 76Y), and Programmer/Analyst (MOS 74F), and the variable race is provide.

Figure 8 below gives the percentage of each race that comprise the given MOS. It is readily seen that the differing MOS's are not comprised of the same proportions of the racial groups. Again this is a general result found with many variables and most MOS's.

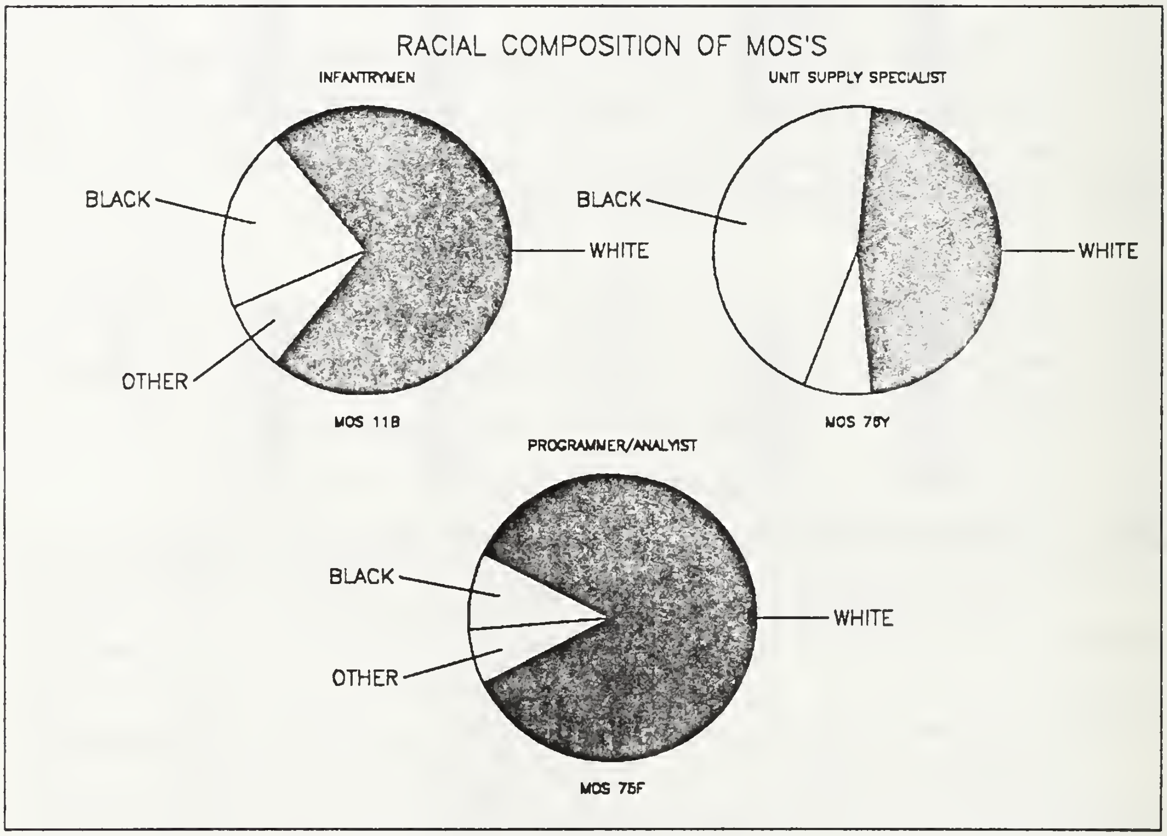

Figure 8. Racial Composition of Three MOS's

The results to this point are as follows:

- MOS's are comprised of categories of soldiers with different probabilities of reenlisting.
- Soldiers in a given category will display similar probabilities of reenlisting in many different MOS's.
- MOS's are comprised of different proportions of the categories.

3. Example of Methodology

Using these observations, we can predict reenlistment rates for MOS's using a procedure illustrated by the following trivial example.

Over the past six years, the reenlistment rate for Infantrymen (MOS 11B) averaged 34%; for the Unit Supply Clerk (MOS 76Y) the rate averaged 46%. An explanation for this difference is that MOS 76Y is comprised of higher proportions of soldiers with higher probabilities of reenlistment. Table 2 provides the example.

Table 2. REENLISTMENT RATES COMPARISONS

Variable	MOS 11B	MOS 76Y	Remarks
Sex	0% Female	21% Female	Females reenlist at a rate 19% higher than males
Race	20% Black	45% Black	Blacks reenlist at a rate 14% higher than whites
Dependent Status	32% Dependents	37% Dependents	Soldiers with dependents reenlist at a rate 20% higher

Again, this trivial example explains the higher reenlistment rate of MOS 76Y by demonstrating that it is comprised of higher proportions of soldiers who reenlist with higher probabilities. This example provides the motivation for our approach.

4. Assumption of the Methodology

A significant assumption is made at this point. The method of this study forms homogeneous groupings of soldiers by looking for similar probabilities of reenlisting. We assume that soldiers with similar probabilities of reenlisting will display similar bonus response rates. Work by one researcher supports this assumption. He shows that soldiers exhibit similar bonus and pay response rates by demographic groups [Ref. 11: p. 212].

5. Motivation for Variable Reduction

There are 40 explanatory variables available to explain the reenlistment decision making process of a soldiers. It is not practical to continue with a 40 dimensional problem, and therefore part of the methodology is to reduce the number of variables. The reasons why this is important are as follows:

- Including 40 variables would require the prediction of those 40 variables each time the model is run.
- Including 40 explanatory variables increases the chance for collinearity within the regression model, which reduces model performance.
- Including 40 explanatory variables (over 20 of which are categorical variables) will require the estimation of over 100 coefficients. A regression equation of this size lacks the parsimony necessary of a good model.
- Most of the explainable variance in reenlistment response rates can be explained with considerable fewer than 40 variables.

Therefore, variable reduction will be an important part of the solution method.

C. METHODOLOGY

As a result of the above discussion, this study adopts the following solution steps.

- Select influential categorical variables using log-linear models.
- Partition the population into cells with similar reenlistment probabilities.
- Reduce the number of cells using cluster analysis.
- Select influential continuous variables using logistic regression.
- Estimate reenlistment rates for each cell using logistic regression.
- Compute projected reenlistment rates for each MOS as a linear combination across all cells.

The use of log-linear models for the categorical variables, and the logistic models for the continuous variables is suggested since the study uses a binary response variable. Influential variables are defined as variables that are likely to be statistically significant predictors of reenlistment rates, and are identified through exploratory data analysis using log-linear and logistic models. The cluster analysis addresses the issue of sparse cells. Cluster analysis, log-linear models and logistic regression are all discussed in more detail in Chapter V, Appendix I and Appendix J.

V. ZONE A ANALYSIS AND RESULTS

A. GENERAL

The purpose of this chapter is to demonstrate the application of the methodology outlined in Chapter IV to the Zone A reenlistment problem.

B. SELECTION OF INFLUENTIAL CATEGORICAL VARIABLES

The first step is to select influential categorical variables, for use in partitioning the Zone A population into cells of soldiers who have similar probabilities of reenlisting.

There are thirty categorical variables available to partition the population, with some of the variables having ten to twenty categories. In the worst case, the problem is partitioned into 8×10^{23} cells. Clearly this is an unmanageable number of cells.

The approach to reducing the number of variables is to use exploratory data analysis techniques. In addition to reducing the number of variables, opportunities to reduce the number of categories within a variable are also explored.

1. Exploratory Data Analysis of Categorical Variables.

This study uses a systematic approach of exploratory data analysis on the categorical variables. It can best be described as a bottom up method. The approach starts by first understanding the data through the study of the variable's distributions and simple univariate procedures, and then increases dimensionality with bivariate and multivariate techniques. This approach is advocated in the data analysis books such as Chambers [Ref. 22: pp. 316-319].

One problem with this approach is that it is impractical to test a large percentage of the interactions of groupings of three or more variables. For example, to test all interactions of three variables would require

$$\binom{30}{3} = 4060 \qquad (2)$$

different models.

Therefore, the study uses an approach outlined in Freeman and Jekel [Ref. 23: pp. 514-519] to discover interesting multivariate groupings. Freeman and Jekel recognize that the variables of potential interest may be hidden in a forbiddingly large cross-classification scheme and that there is a tradeoff between trying to reduce the number of variables and the potential of losing valuable information. Therefore, they propose the following procedure.

- Perform a test for independence between each pair of variables.
- If two variables are dependent, then form a compound variable using them. Compound variables are formed by combining two variables together into a single variable with categories corresponding to all combinations of categories of the variables being combined.
- Perform a test for independence between these compound variables and all other variables.
- Form new compound variables for each pair consisting of a compound variable and a single variable that are dependent.
- Continue this process until cell frequencies becomes small (less than one.) At this point, terminate the selection process, and choose the variables with the most significant associations for inclusion in the reduced table.[21] [Ref. 23: pp. 513-518]

The goal of this section is to produce a parsimonious model [Ref. 24: p. 156]. For reasons of readability, we do not present every test conducted within the paper. Instead an example or two is presented to show the procedure, and than the results summarized.

2. Exploratory Data Analysis Tools

There are two primary type models to use on categorical data. They are linear models, as described by Grizzel, Starmer and Koch [Ref. 25: pp. 491-492] and log-linear models, as described by Bishop, Fienberg and Holland [Ref. 26: pp. 28-37].

This study will primarily use the log-linear models for the study of categorical variables. Log-linear models work especially well in analyzing contingency tables of three or more dimensions [Ref. 27: p. 207] and are useful in testing hypotheses about the nature of relationships between two or more categorical variables [Ref. 24: p. 143]. Appendix H gives the background of log-linear models.

3. Distribution of Variables

The first step in the systematic approach to data analysis is to study the distributions of the individual variables. Table 3 lists the thirty categorical variables, and gives the range and type of measurement scale of the variable. The right column is explained below.

---

[21] The procedure outlined does not guarantee selection of the best table, nor should it always be followed rigorously. Instead in the spirit of exploratory data analysis, it is a rational, easily implemented procedure to select an interesting table.

Table 3. MEASUREMENT SCALE AND RANGES FOR CATEGORICAL VARIABLES

Variable Name	Range of Values	Measurement Scale	Skewed
ACF	0-8	Nominal	Yes
Enlistment Bonus	0-6	Nominal	Yes
Enlistment Term	2-4	Ordinal	Yes
Enlistment Program	1-21	Nominal	No
Age at Enlistment	17-34	Interval	No
Age at Separation	19-40	Interval	No
Prior Service	0-6	Nominal	Yes
Reserve Time	0-1	Nominal	Yes
Youth Program	0-7	Nominal	Yes
Hometown (Region)	0-10	Nominal	No
Education at Enlistment	1-12	Ordinal	Yes
Education at Reenlistment	1-12	Ordinal	Yes
Change in Education	0-1	Nominal	Yes
Dependent Status at Enlistment	10-29	Nominal	Yes
Dependents at Reenlistment	10-29	Nominal	Yes
Change in Dependents	0-1	Nominal	Yes
Character of Service	0-1	Nominal	Yes
Mental Test Category	1-8	Ordinal	No
Years of Service	2-6	Interval	No
Current Rank	1-6	Ordinal	Yes
Duty Location	1-13	Nominal	No
Race	1-3	Nominal	Yes
Ethnic Group	1-6	Nominal	Yes
Sex	1-2	Nominal	Yes
Job Type	0-9	Nominal	No
Retirement System	0-1	Nominal	Yes
Number of Years to Military Retirement	2-20	Interval	No
Type of Bonus Payment	1-2	Nominal	Yes
Job Skill Migration	1-2	Nominal	Yes
Reenlistment Bonus Multiplier	0-6	Interval	Yes

The most significant result of the study of individual distributions concerns the number of observations in each category. Variables are of two types. One type, of which the variables **TERM OF ENLISTMENT** and **SEX** are typical, have a large number of observations in one category. Figure 9 shows the uneven frequency distribution of **TERM OF ENLISTMENT** and **SEX**. Table 3 has a *Yes* in the right column for variables of this type.

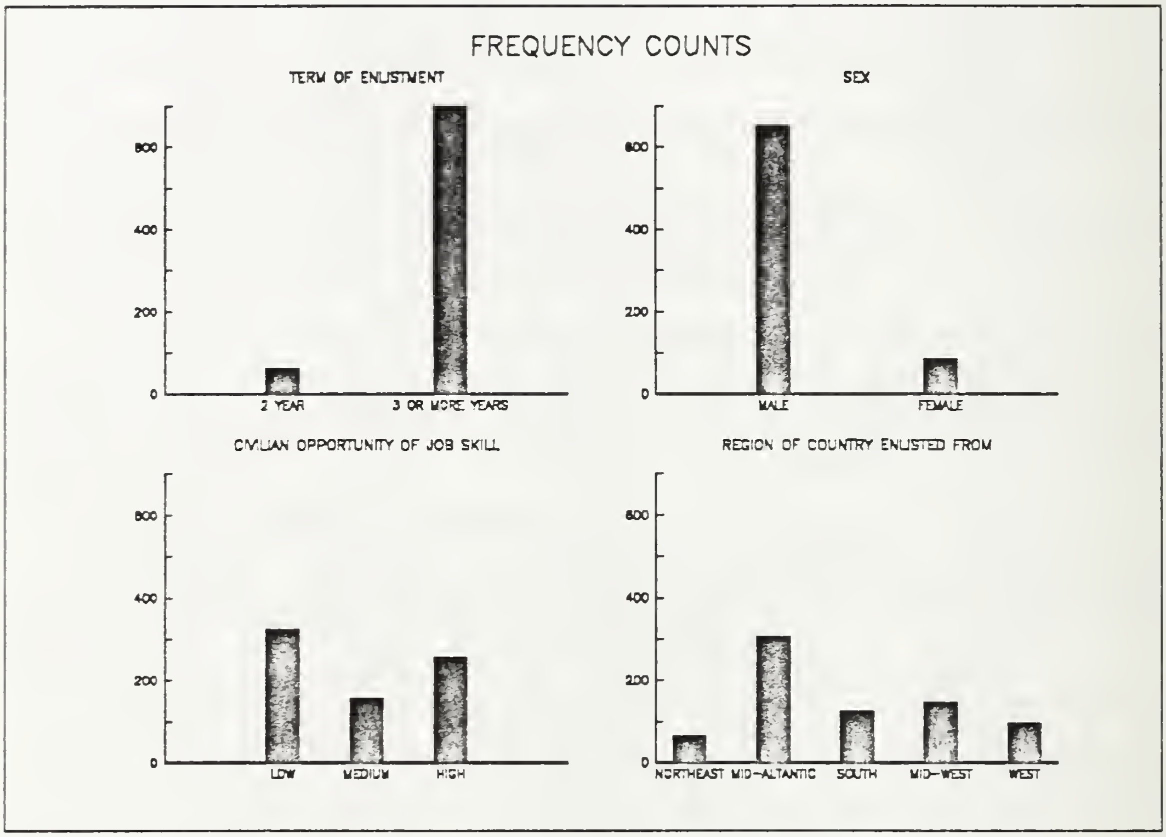

Figure 9. Frequency Counts For Selected MOS's

The second type variable, of which **CIVILIAN OPPORTUNITY OF JOB SKILL** and **REGION OF COUNTRY ENLISTED FROM** are typical, have the bulk of frequencies spread over many values. Figure 9 shows the larger number of categories with a significant number of observations for the variables **CIVILIAN OPPORTUNITY OF JOB SKILL** and **REGION OF COUNTRY ENLISTED FROM**. These variables have a *No* in the right column of Table 3.

When the population is partitioned using variables that have a large number of observations in one category (and therefore other categories with extremely small number of observations), this causes a large number of sparse cells. The issue of sparse cells is addressed in great length later in the study; however, it is important to understand the causes of those sparse cells.

4. Univariate Analysis

The first result of univariate analysis concerns variables having interval measurement scales. Figure 10 shows the reenlistment rates for the categorical variable **AGE AT ENLISTMENT**, an example of a variable with an interval measurement scale. Clearly the older soldiers are, the higher their probability of reenlisting. However, the variance increases significantly as age increases, due to the decreasing number of observations.

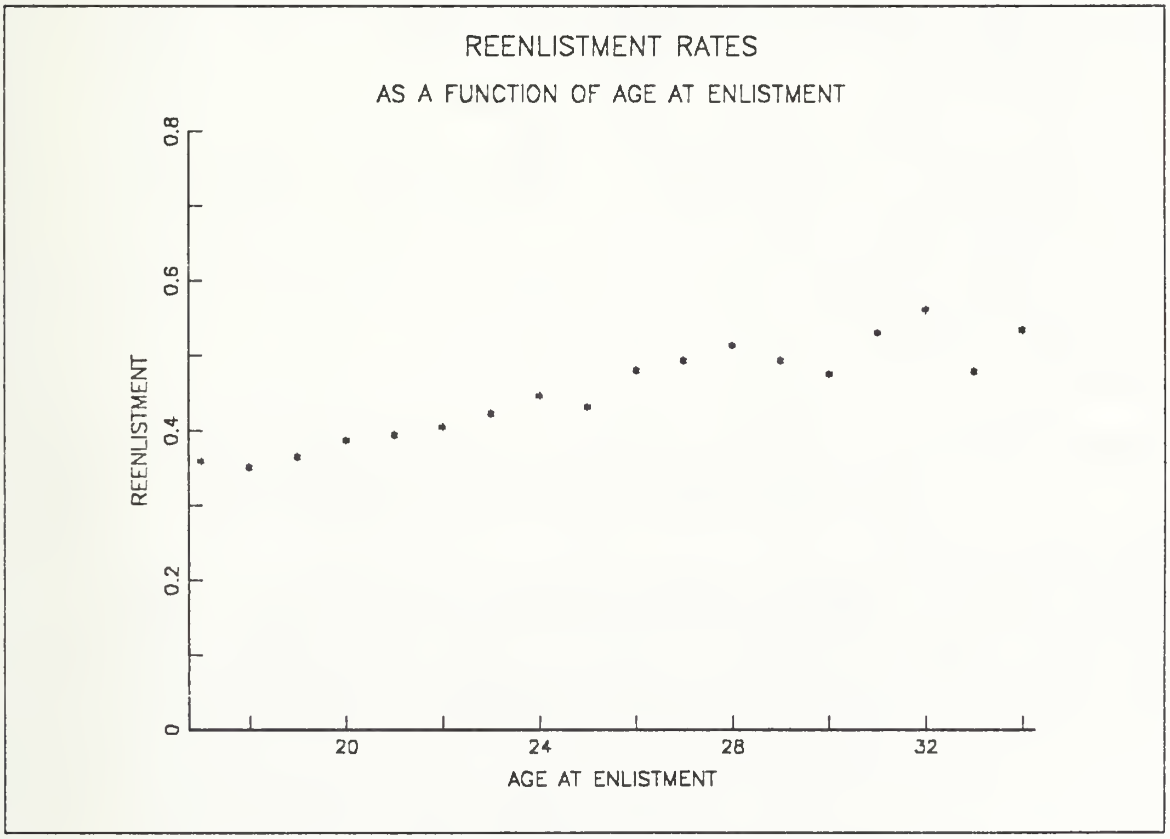

Figure 10. Reenlistment Rates for all MOS's, by Age at Enlistment

AGE AT ENLISTMENT is one of the interval variables that can be treated either as a categorical variable or as a continuous variable. Although it could recoded into fewer categories, it is not intuitive to do so, because of the generally increasing probability to reenlist as age increases. Additionally, because the bulk of the observations are in the left tail, numerous sparse cells result. Analysis such as this leads us to drop the following variables from consideration as categorical variables. They will be reconsidered as continuous variables.

- Age at Enlistment
- Age at Separation
- Years of Service
- Number of Years to Military Retirement
- Reenlistment Bonus Multiplier

There are numerous variables in which hypothesized relationships are not validated by the univariate analysis. Among these are:

- Enlistment Bonus
- Enlistment Program
- Youth Program
- Retirement System
- Type of Bonus Payment
- Job Skill Migration
- Reserve Time
- Duty Location

Some of these variables are rejected due to data problems. For example, **ENLISTMENT BONUS** has far fewer number of soldiers coded as receiving a reenlistment bonus then are known to have received them. Some of the variables are dropped because there is no significant difference in the reenlistment probabilities for different categories. For example, **ENLISTMENT PROGRAM** is dropped for this reason. Finally, some variables are discarded because of interactions with other factors. For example, **DUTY LOCATION** is discarded because analysis shows reenlistment rates of over 95% for soldiers stationed overseas. However, further analysis shows that soldiers who near the end their term of service overseas are brought back from overseas prior to their discharge, while reenlisting soldiers remain overseas. If not corrected for, this leads to a biased assessment of the effect of **DUTY LOCATION** on the reenlistment rate.

The final univariate analysis result involves reduction in the number of categories in certain variables. Figure 11 shows why **MENTAL CATEGORIES** are recoded from seven categories to four categories. Categories 2-5 have statistically similar reenlistment probabilities, and therefore are recoded into one category.

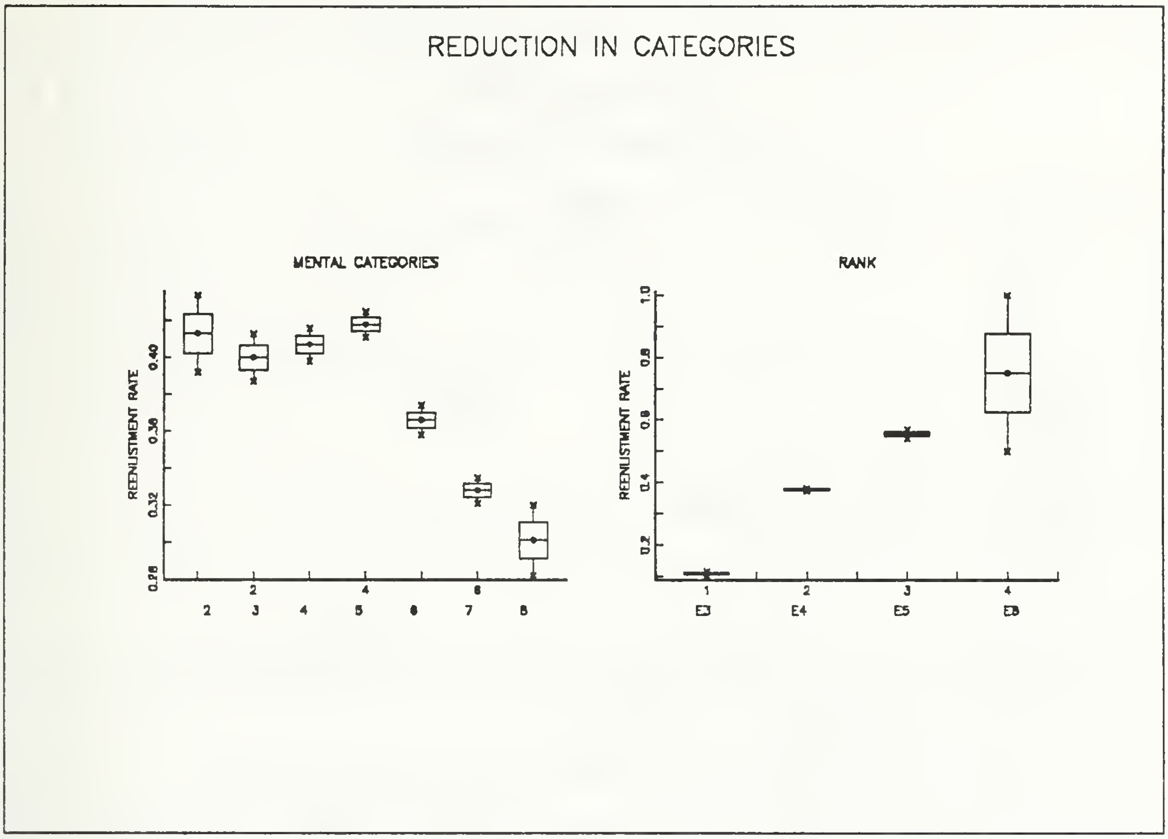

Figure 11. Reenlistment Rates by Mental Category and by Rank

Figure 11 shows how the variable **CURRENT RANK** is recoded as three groupings, even though there clearly appear to be four distinct groupings. However, when the frequency numbers are examined, the E6 category contains less than 200 of the 75,788 observations. Since the E6 category is not statistically different from the E5 category, they are combined without loss of precision.

Analysis shows significant differences in reenlistment rates by home state. Clearly, however, including the fifty state categories is impossible. Since, there appear to be regional trends, the first step is to categorize the states into the nine standard United States regions. While categorization into these regions is a good first step, there are still some inconsistencies, and the number of categories is still too great. Therefore, the states are further categorized into five regions. Figure 12 shows the reenlistment rates for those five regions. Analysis shows that these categories are stable over time.

Similarly, the Army's 350 military job specialities are grouped into three general categories, which is our subjective evaluation of the civilian opportunities available to soldiers with different job skills.

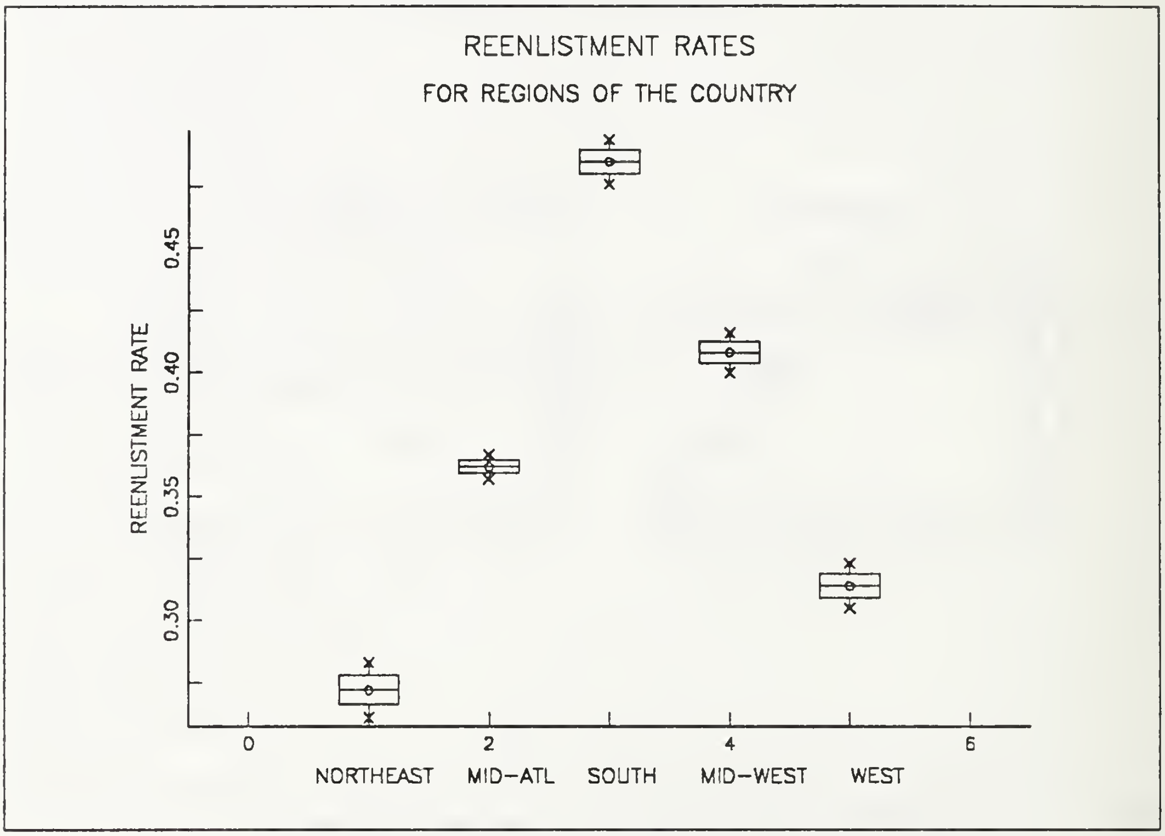

Figure 12. Reenlistment Rates for Regions of the Country

At the end of the univariate analysis, 17 variables remain. All have between two and five categories.

5. Multivariate Analysis

One of the purposes of the multivariate analysis is to choose between groups of variables that are clearly collinear. The first of these groups are the variables which measure education levels.

- Education at Enlistment
- Education at Reenlistment
- Change in Education

The second group measures dependent status.

- Dependent Status at Enlistment

- Dependent Statue at Reenlistment
- Change in Dependent Status

The third group measures race and ethnic groups.

- Race
- Ethnic Group

The analysis confirms the dependence between the variables, and gives guidance as to the best variables to select. The variables are:

- Education at Reenlistment
- Dependent Statue at Reenlistment
- Ethnic Group

As a result of this analysis, 12 categorical variables are retained. These 12 are listed in Table 4, along with their final categories.

Table 4. REMAINING CATEGORICAL VARIABLES

Variable Name	Range of Values	Measurement Scale	Symbol
ACF	0-1	Nominal	C
Enlistment Term	2-3	Ordinal	T
Prior Service	0-1	Nominal	P
Hometown (Region)	1-5	Nominal	H
Education at Enlistment	1-3	Ordinal	E
Dependents at Reenlistment	1-2	Nominal	D
Character of Service	0-1	Nominal	X
Mental Test Category	5-8	Ordinal	M
Current Rank	3-5	Ordinal	G
Race	1-3	Nominal	R
Sex	1-2	Nominal	S
Job Type	1-3	Nominal	J

6. Table Selection

To further reduce the number of variables, the procedure (described on page 32) by Freeman and Jekel [Ref. 23: pp. 514-519] is applied to the remaining 12 variable. The first step in selecting the multi-dimensional table is to examine the dependence of all pairs of variables. The analysis of the dependence uses Cramer's test [Ref. 23: pp. 514-519] as a measure of association. The significant pairs of variables are TD GR SR RH and JE. This first table is not displayed due to its size, however it is constructed similar to Table 5 below.

The second step in selecting the multi-dimensional table is to form a compound variable from each dependent pair of variables as described on page 32, and then test the dependence of the compound variables with all remaining variables [Ref. 23: p. 517]. Table 5 shows the results.

Table 5. ASSOCIATIONS WITH COMPOUND VARIABLES

Variables		TD	GR	SR	RH	JE
	Levels	4	9	6	15	9
C-ACF	2					
T-Enlistment Term	2					
P-Prior Service	2					
H-Hometown (Region)	5		X			
E-Education at Enlistment	3					
D-Dependents at Reenlistment	2					
X-Character of Service	2					
M-Mental Test Category	4					
G-Current Rank	3	X				
R-Race	3					X
S-Sex	2					
J-Job Type	3			X		

Significant tables are TDG, SRJ, JER, and HGR. Continuing on in this manner leads to the following results.

7. Results of Exploratory Data Analysis

As a result of the exploratory data analysis, the following variables are used to partition the data set:

Term	(2 categories)
Rank	(3 categories)
Sex	(2 categories)
Race	(3 categories)

Dependents	(2 categories)
Region	(5 categories)
Job Type	(3 categories)

C. PARTITIONING OF THE POPULATION INTO HOMOGENEOUS CELLS

The purpose of this step is to partition the population into homogeneous cells containing soldiers with similar probabilities of reenlisting. The variables are the influential categorical selected in the above step.

Using the seven categorical variables with between two and five categories each to partition the population creates a total of 1080 cells. A random sample of 75,788 Zone A soldiers shows that 859 of the cells have non-zero frequencies, 162 over 100 observations, and 12 over 1000 observations.

Clearly, this is too many cells. Additionally, the sparse cells (those approximately 550 cells with under 25 observations) do not perform well in regression. Therefore, further reduction of the number of cells must occur.

D. CELL REDUCTION

1. Cell Reduction Procedure

There is considerable literature concerning cell reduction of multidimensional contingency tables. These studies identify three primary ways to reduce multidimensional tables [Ref. 28: p. 546] [Ref. 29: pp. 328-329]. These three methods are:

- Reduce the Number of Variables
- Reduce the Number of Categories in a Variable
- Combine Cells Within the Multidimensional Contingency Table

Of these three techniques, the first two are fully exploited in previous sections. Analysis shows that further reduction using these techniques results in significant loss of information. Therefore, we turn to techniques to combine cells within the multidimensional table to further reduce the number of cells.

Combining cells within the multidimensional table using cluster analysis is the technique used in a thesis by Larsen [Ref. 30: pp. 22-34]. The problem he solves is estimating retention rates for Marine Corps officers. He partitions his population into cells using years of service, job speciality, and source of commission. Similarly to this thesis, he ends up with many sparse cells, and combines them using cluster analysis.

While this study does not use the computerized cluster analysis techniques of the Larsen study, the ad-hoc procedure used follows the same principles. The primary reason for not using the computer package is the existence of special structure in the problem, which is not fully exploited by the package.

The special structure in this problem is the existence of a subset of variables which have a large percentage of the observations in one category, and therefore other categories with few observations. An example of this is the variable **SEX**, which has less then 8% woman. An extremely large proportion of the cells that have this category associated with it are sparse cells.

The second part of the special structure is that the variables having the large percentage of the observations in one category also have the most significant differences in probabilities to reenlist between cells. For example, in the case of the variable **SEX**, the category **WOMEN** is a relatively homogeneous grouping, requiring little further categorization. The ad-hoc procedure of this study exploits this structure to combine cells by examining the variables in the following order:

- Term of Enlistment
- Sex
- Rank
- Dependents
- Race
- Region
- Job Type

This ordering examines those variables with the largest percentage of large categories first.

2. Cell Reduction Results

Using the ad-hoc cluster analysis procedure reduces the number of cells from 1080 to 92. All cells have at least 37 observations (from a random sample of 75778 observations). Only five of the cells have under 100 observations, and 24 of the cells have over 1000 observations.

Although variable reduction is proceeding, there are still too many cells. Therefore cells are further combined, this time by grouping cells with similar reenlistment probabilities. Cells are grouped only if they fall into a three percentage point window. Attempts are made to group like cells; this goal is slightly relaxed to facilitate groupings.

36 cells result from the second iteration of cell reduction. Reenlistment rates vary from 7% to 80% within these cells. The smallest cell has 232 observations from a

75778 observation sample, and 20 of the 36 have over 1000 observations. Appendix J lists the composition of each of the 36 cells, and the reenlistment rates for each group.

E. SELECTION OF INFLUENTIAL CONTINUOUS VARIABLES

1. Exploratory Data Analysis of Continuous Variables

The purpose of this section is to select the influential continuous variables for inclusion in the regression equations. The technique is exploratory data analysis, using a bottom up approach as described earlier in this chapter. The primary tool is logistic regression. Appendix I describes these techniques in detail.

The section begins with 20 potential variables. The goal is to choose five to seven for inclusion in the regression equations.

Since the reenlistment population is partitioned into 36 different cells, this analysis could be preformed separately for each cell. However, this entails a prohibitive amount of work. Instead the exploratory data analysis is performed on the entire population. This is compensated for by the separate stepwise regression on each cell.

A general observation of the exploratory data analysis is that although there are significant relationships between many of the explanatory variables and the response variable, few of the variables account for a large portion of the variance in reenlistment probabilities. This result lowers considerably the expectations for the amount of the variance the overall model explains.

2. Distribution of Individual Variables

The purpose of this section is to examine the distribution of the continuous variables. The logistic regression model requires no specific distributional assumptions (for example normality). However, the regression model gives inaccurate estimates if the variables do not have sufficient range and spread. Table 6 shows the range, mean, and standard deviation for the continuous variables. All the variables have adequate range and spread. A second issue is the scale of the variables in relationship to each other. Regression techniques often do not perform well if the variables are widely scaled. The scales in this case are moderate, and a well-behaved model is anticipated.

3. Univariate Analysis

The primary purpose of the univariate analysis is to select the influential variables for inclusion in the regression equations.

Figure 13 gives the results of a logistic regression to test the significance of the variable **BONUS LEVEL** on the probability of reenlisting, using the SAS LOGIST

Table 6. RANGES, MEANS AND STANDARD DEVIATIONS FOR CONTINUOUS VARIABLES

Variable Name	Range of Values	Mean	Standard Deviation
Unemployment Rate at Enlistment	2.4, 18	7.75	2.33
Unemployment Rate at Reenlistment	2.4, 18	7.81	2.39
Promotion Rates	-38, 95.5	-0.18	7.31
AFQT Score	0, 99	49.89	23.38
Age at Enlistment	17-34	19.65	2.59
Age at Separation	19-40	22.88	2.73
Consumer Price Index	1.1, 8.9	3.73	1.36
Gross National Product	0.037, 0.117	0.070	0.020
Years of Service	2, 6	3.87	0.78
Number of Years to Military Retirement	14, 18	16.13	0.78
Real Military Compensation	2, 12	4.36	2.93
Promotion Rate Forecast	-38, 95.5	-0.18	7.31
Reenlistment System	1, 5	2.81	1.35
Bonus Multiplier	0, 5	0.49	0.89
Real Military Compensation (Inflation Adjusted)	2, 12	4.36	2.93

procedure. Of note are two items. First is the low R value. Appendix I discusses the R value for logistic regression in detail; it is analogous to the R in ordinary least square regression, which is a measure of the fit of the model. The second item of note is the p value. This represents the following hypothesis test.

$$H_0: \textit{Coefficient Estimate is Zero} \tag{3}$$

$$H_1: \textit{Coefficient Estimate is Not Zero} \tag{4}$$

The specific test is a Wald test for zero slope, and the test statistic is closely approximated by a Chi-square distribution [Ref. 31: p. 191]. The low p value in Figure 13 represents a low probability that the variable **BONUS** has a slope of zero, and therefore a low p (< 0.05) represents the rejection of the null hypothesis, and strongly suggests that the bonus does have a effect on reenlistment rates.

```
                LOGISTIC REGRESSION PROCEDURE


                  DEPENDENT VARIABLE: RCODE


73481 OBSERVATIONS
45697  LEAVE    =  0
27784  REUP     =  1
          0 OBSERVATIONS DELETED DUE TO MISSING VALUES

VARIABLE        MEAN     MINIMUM      MAXIMUM       S. D.

BONUS       0.485935           0            5     0.88916



CONVERGENCE IN 15 ITERATIONS       R= 0.060.



VARIABLE    BETA   STD. ERROR  CHI-SQUARE    P       R

INTERCEPT -0.576    0.0087      4349.01    0.001
BONUS      0.158    0.0084       354.48    0.001   0.060
```

Figure 13. Regression of Bonus Level vs Reenlistment Probability

The above example has an estimation of the intercept term of -0.576 and a slope of 0.158 for the variable **BONUS LEVEL**. These, however, are the transformed intercepts (see Appendix I for a full explanation). To get the actual reenlistment probability at a given bonus level Equation 5 is used, where α and β are the intercept and slope terms, and X is the bonus level.

$$P_i = \frac{1}{1 + e^{-\alpha - X\beta}} \tag{5}$$

A plot of this function is in Figure 14.

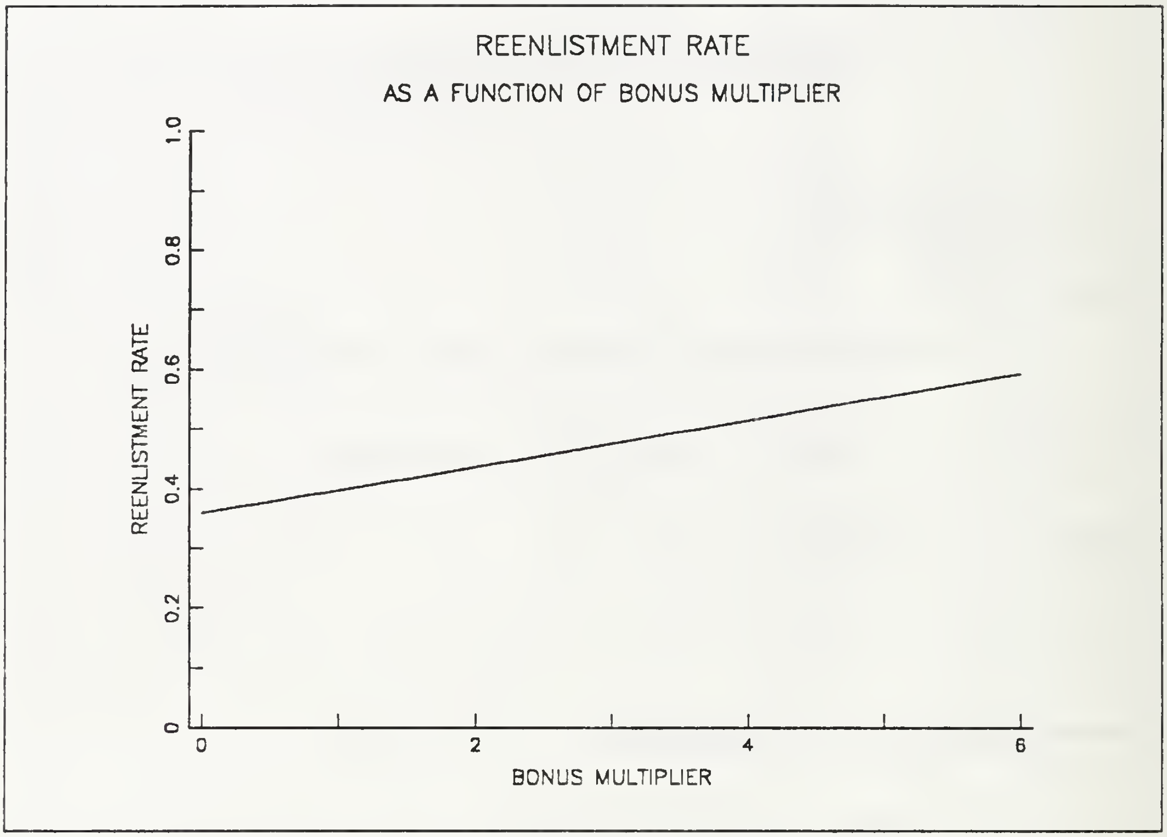

Figure 14. Plot of Bonus Level vs Reenlistment Probability

A second purpose of the univariate analysis is to "fine tune" the variables. An example of this is to plot the unemployment rate just prior to a soldiers reenlistment date, and also lagged by two months, then six months and nine months, and see which is most influential on the reenlistment probability. The issue is much more complicated than this however, because there are issues of which unemployment rates to choose (for the entire population or for certain age groups), whether to choose local regional or national rates, and whether to choose unadjusted or seasonally adjusted rates. Clearly this level of detail is beyond the scope of this thesis; whole studies have addressed just the one issue of which unemployment rate to use. Some limited work is done on the continuous variables; however, for the most part we have relied on the literature to point the way in choosing continuous variables. The limited results achieved in this analysis are incorporated in Chapter III.

4. Bivariate and Multivariate Analysis

One major issue of this analysis is collinearity. When variables included in the regression are collinear or linear combinations of each other, they reduce the precision of the coefficient estimates. There is significant potential for collinearity in the estimation of reenlistment rates. The reason is that longer soldiers remain in the service, the higher their probability of reenlistment becomes. Therefore, any variable that increases as a function of a soldier's time in the service shows a positive correlation with the reenlistment probability. Examples of these variables are many. Rank increases with a soldier's increasing time in service, and pay amount is a function of rank and time in the service. Generally the number of dependents a soldier has increases with service, as does his education level, and his age. A soldier's initial term of service is positively correlated with his time in service. These are all examples of potentially collinear variables, which may adversely affect the precision of the coefficient estimates. Therefore, extreme care is taken to ensure that variables that are collinear are not included.

To test for collinearity, regressions are performed on pairs of potentially collinear variables. If the variables display a high R value, then they are highly collinear, and one of the variables is not included in the regression model. For example, the two variables, **AGE AT ENLISTMENT** and **AGE AT SEPARATION** are potentially collinear. A regression of these variables has an R value of 0.9229. This high R value is the first clue of the collinearity of these variables. If collinear variables are included, the regression model will indicate a better model fit than is justified by the data. A full explanation of collinearity, and its effects on regression models is found in Mosteller and Tukey [Ref. 32: pp. 280-284].

5. Results of Exploratory Data Analysis

As a result of the exploratory data analysis of the continuous variables, the study includes the following variables in the regression models:

- Unemployment Rate at Reenlistment
- Promotion Rate
- AFQT Score
- Pay
- Bonus Level
- Reenlistment System

- Age at Entry

F. ESTIMATION OF REENLISTMENT RATES

A stepwise logistic regression is performed on each of the 36 cells, using the procedures outlined in Appendix I. Appendix K contains a table of results. The table contains the estimated coefficients, plus the R value for each regression. Additionally Appendix K gives the results of the hypothesis test to see if the coefficient is statistically different from zero.

Equation 6 below gives a example of the bonus equations for one of the cells, Cell 22.

$$P_i = \frac{1}{1 + e^{1.09 \; - \; 0.209 \times Bonus \; + \; 0.012 \times AFQT \; + \; 0.057 \times Age\ at\ Entry}} \qquad (6)$$

Analysis of the results in Appendix K leads to the following observations:

- The R values for all the regression equations are low. This was expected, as the estimation of reenlistment rates is a difficult problem. This is because many factors play into a soldiers decision to reenlist; we can only hope to capture some of those reasons with measurable variables.
- Although the R values are small, the explanatory variables included have low p values, indicating that the slope of the estimated coefficient is significantly different than zero.
- There are some cells for which the bonus level did not significantly influence the reenlistment rate.

G. COMPUTATION OF MOS REENLISTMENT RATES

The final step to the procedure is to calculate the reenlistment rate for the MOS, as a linear combination across all the cells. To illustrate how this is done, an example is provided.

In this example, the reenlistment rates for MOS 11B (Infantryman) are computed for 1990. The following information is estimated for next year.

- The unemployment rate will be 5.0%.
- MOS 11B's promotion rate average will be higher than other MOS's, so that the average 11B soldier is promoted six months sooner than the average.
- The AFQT average score will be 63.
- The pay raise for next year will be 3.2%
- The reenlistment system will remain liberal
- Additionally, the average 11B soldier eligible to reenlist next year was 19 years old when he enlisted.

Figure 15 gives the projected breakdown, by cell, of MOS 11B for soldiers eligible to reenlist next year. Computing the reenlistment rate for MOS 11B gives the results in Table 7.

Table 7. REENLISTMENT RATES FOR MOS 11B

Bonus Level	Reenlistment Probability
0.0	23.7%
0.5	29.1%
1.0	35.1%
1.5	41.6%
2.0	48.5%
3.0	62.1%

H. MODEL VALIDATION

Since the data set was partitioned prior to the beginning of the analysis, cross-validation of the regression models is possible using the remaining data.

The cross-validation is conducted on the 36, rather than on the 350 MOS's. Table 8 shows the results of a randomly selected number of the cells. The first column shows the estimated reenlistment rates for the cell over the past six years. The second column has the actual reenlistment rates. The excellent fit of the model is seen just by comparing these two columns. The fit is confirmed through use of a chi-square goodness-of-fit test. The procedure followed is the same as described in Appendix J. The model is rejected at the $\alpha = 0.05$ level, if the test statistic is greater than 3.841. Clearly, these result confirm the validity of the regression models.

A second part of the model validation is to check the residuals of the regression model. There are no indications of problems with the residuals. Appendix I discusses the form of the logistic regression residuals.

I. MODEL PRECISION

The military reenlistment bonus model is a deterministic model which optimizes estimated means, and requires point estimates of reenlistment rates. However, we feel obligated to discuss confidence intervals on those point estimates. We recommend the that the users of the military reenlistment bonus model conduct sensitivity analysis, by varying reenlistment rates in order to understand how the estimate impacts on their decisions. The confidence intervals provide guidance on the reenlistment rate values that should be used for worst and best case estimates.

Table 8. **RESULTS OF MODEL VALIDATION**

Cell Number	Estimated Reenlistment Rate	Actual Reenlistment Rate	Error	T Statistic
Cell 1	30.5%	31.3%	+0.8%	0.27
Cell 2	24.2%	25.1%	+0.9%	0.42
Cell 7	27.3%	24.3%	-3.0%	1.88
Cell 12	48.6%	45.3%	-3.3%	1.97
Cell 22	36.4%	37.1%	+0.7%	0.38
Cell 24	40.3%	38.4%	-1.9%	1.71
Cell 43	61.4%	58.5%	-2.9%	0.80
Cell 47	40.8%	43.5%	+2.7%	1.38

The military reenlistment bonus model does not accept confidence intervals as model inputs. Therefore, instead of generating a table of 350 MOS confidence intervals that would not be used, we instead provide a general rule of thumb to guide the selection of values for sensitivity analysis. Generally, the predicted rate + - 10% gives a 70% confidence interval, the predicted rate + - 15% gives a 95% confidence interval. These worst case estimates also attempt to account for additional error that results from inaccuracies in estimating the inputs to the reenlistment model, such as the unemployment rate.

CELL	NUMBER	PERCENT	CUMULATIVE NUMBER	CUMULATIVE PERCENT
1	107	1.6	107	1.6
2	35	0.5	142	2.2
3	610	9.3	752	11.5
5	36	0.5	788	12.0
6	390	5.9	1178	18.0
7	21	0.3	1199	18.3
8	9	0.1	1208	18.4
22	304	4.6	1512	23.1
24	223	3.4	1735	26.5
26	716	10.9	2451	37.4
28	437	6.7	2888	44.0
31	230	3.5	3118	47.6
37	93	1.4	3211	49.0
38	137	2.1	3348	51.1
39	6	0.1	3354	51.2
41	52	0.8	3406	51.9
46	983	15.0	4389	66.9
49	90	1.4	4479	68.3
51	75	1.1	4554	69.5
52	131	2.0	4685	71.5
58	98	1.5	4783	72.9
63	177	2.7	4960	75.6
66	228	3.5	5188	79.1
72	8	0.1	5196	79.2
73	1118	17.1	6314	96.3
76	243	3.7	6557	100.0

Figure 15. Breakdown of MOS 11B by Cell

VI. CONCLUSIONS

A. FINDINGS

This study develops a methodology for estimating reenlistment rates for use in the military reenlistment bonus model. It departs significantly from methods of previous studies in that it does not group MOS's into skill families or other similar groupings. Instead this study looks for homogeneous groupings of soldiers with similar probabilities of reenlisting, and develops regression models for these groupings.

There is strong statistical evidence that certain groups of soldiers have very different reenlistment propensities. These groupings are best defined by categorical variables, which partition the population into cells of soldiers who are homogeneous with respect to their reenlistment probability. This study assumes that these groups are also homogeneous with respect to their response to changes in bonus levels. There is some prior research to support this assumption [Ref. 11: p. 212].

Many researchers include one or two categorical variables in their regression equations. Few, however, exploit the full potential of these variables. Including more categorical variables leads to many cells with low expected frequencies.

To overcome the low expected frequencies, this study first partitions the population into cells and then groups cells. The grouping procedure uses the principles of cluster analysis to take advantage of special problem structure by finding the variables most likely to create low expected frequency cells. The resulting grouped cells contain soldiers with nearly the same statistical reenlistment probabilities. Regression models are developed for each grouping of cells, and MOS reenlistment rates as a function of bonus level are calculated as a linear combination across the cells.

Most of the regression equations had low R^2 values. These low R^2 do not invalidate the model for several reasons. First, the grouping of the cells by clustering is a variance-reduction step. The R^2 for the regression models indicate the amount of variance within the groups that is explained. Since the grouping of cells reduces the variance within a cell, the potential for further reduction is limited. Second, while the R^2 is low, the variables included in the regression models are statistically significant. Third, the study is hampered by the quality of the national economic variables. Variables such as **GNP**, **UNEMPLOYMENT RATE** and **CIVILIAN JOB GROWTH** are quantified at an aggregated level. Finer resolution data (by quarter and by geographic location) would help further explain variance. Fourth, the low R^2 value is not unexpected in this type of problem. This study tries to explain a soldier's reenlistment propensity using nationally measurable variables. However surveys of soldiers show that the reenlistment decision making process is complex, involving issues as complex (and unmeasurable) as a soldiers relationship with his peers, and his job satisfaction. Given this, it is not surprising that the R^2 is low. Finally, despite the low R^2, the models are validated using cross-validation. This cross-validation finds the models to be a highly predictive, credible models of significant value.

A noteworthy finding of this study is that the variable **BONUS LEVEL** is not significant in numerous cells. In other words, soldiers in these cells do not respond to increasing cash bonuses. Obviously bonuses should not be allocated to MOS's with high percentages of soldiers from these cells.

One of the difficulties of this study is the inability to quantitatively measure items such as the effectiveness of the reenlistment system in providing soldiers with their desired reenlistment option. However, the results of the subjective variable **REENLISTMENT SYSTEM** are extremely interesting. This variable measures how "liberal" the reenlistment system is in providing soldiers their reenlistment options. It is significant in as many equations as is the bonus level. The most recent improvement in this area is a program called the *Commander's Override*, in which the computerized reenlistment system is manually overriden to keep a soldier in the service by providing his or her reenlistment option choice. Clearly programs such as these are an alternatives to the cash reenlistment bonus.

Another finding is the significance of the variables to measure a soldier's motivation to join the service. These enlistment variables are important in determining the first term reenlistment model. Among these variables are **TERM OF ENLISTMENT, SEX, RACE, REGION, JOB TYPE** and **AFQT PERCENT**. Since many of the enlistment variables are significant in the Zone A reenlistment model, further study of other enlistment variables is in order. There is an enlistment data base which was not available for this study that contains numerous variables of potential interest. Since enlistment demographics appear significant to the first-term reenlistment decision, then one way to improve first-term reenlistments is to target for enlistment those groups of soldiers who display the highest reenlistment propensities.

A finding of this study is that the potentially complicating issues of MOS migrations, extensions and reenlistment windows can be ignored, with only minor loss of accuracy in the reenlistment estimates. This greatly simplifies the reenlistment model. Appendix B discusses this issue in detail.

This study developed a alternative technique to previous methods of grouping MOS's. This method was cross-validated with data not used in the model development of the model. The results are highly predictive of reenlistment rates, and responses to bonuses.

B. RECOMMENDATIONS

The estimates of Zone A reenlistment rates developed in this study should be adopted for use in the military reenlistment bonus model.

The procedures outlined in this study should be replicated to estimate the Zone B and Zone C reenlistment rates.

C. RECOMMENDATIONS FOR FURTHER STUDY

- This study does not analyze the composition of the grouped cells to any great extent. However, one could potentially gain considerable insight into the reenlistment decision making process from exploring the composition of each cell, and explaining why certain groups of soldiers cluster together. Similarly, detailed examination of the cells in which the bonus level is significant should be conducted in order to understand what types of soldiers respond to bonuses and why.
- Further attempts need to be made to quantify and study the force alignment variables (such as pay, promotion rates and the form of the reenlistment system) which impact on the reenlistment program. These variables are potentially as powerful as the reenlistment cash bonus.
- The enlistment data base from the Military Entrance Processing Command should be examined for further enlistment variables to explain the first term reenlistment decision. This data base was not available for this study. Several enlistment variables were significant in this study's model, however, there are many other enlistment variables still to examine. Examples of variables that should be examined include variables that measure a the income of a soldier's parents and the military background of the soldiers parents and siblings.
- This study used a type of cluster analysis procedure to reduce the number of cells. However, numerous other techniques are available for use. Many of the techniques are discussed in a thesis by Misiewicz [Ref. 33: pp. 1-15]. Further research should examine these additional procedures, particularly shrinkage using Empirical Bayes.
- The annualized cost of leaving (ACOL) model described in Chapter II, together with more detailed economic variables should be incorporated into this methodology.
- Finally, as an alternate solution technique, the use of intervention analysis should be explored. An article by Box and Tiao should serve as a starting point. [Ref. 34: p. 70].

APPENDIX A. THE MILITARY REENLISTMENT BONUS MODEL

A. GENERAL

The military reenlistment bonus model is a mathematical programming model for optimizing the allocation of reenlistment cash bonuses in order to achieve the desired force structure. The model is essentially a deterministic model. The model was developed at the Naval Postgraduate School by Major Dean DeWolf, Major Jim Stevens, and Professor Kevin Wood, and is currently used by the U. S. Marine Corps and the U. S. Army [Ref. 1: pp. 1-3].

B. INPUTS

The inputs for the model are by military occupation speciality (MOS). They include:

- Current force structure
- Desired force structure
- Number of soldiers eligible to reenlist
- Training costs
- *Projected* reenlistment rates at each bonus level[22]

Additionally, inputs include the bonus budget, and the maximum size bonus a soldier is eligible to receive.

C. OUTPUT

The output from the model is recommended bonus levels for each of the 350 MOS's in each of their three zones. The model also outputs the projected force structure after the bonus payments.

D. OBJECTIVE FUNCTION

The objective function measures the deviation from the desired force structure. Deviations in some MOS's are weighted higher because of the MOS's criticality, or because of the higher investment in training the Army has in certain soldiers.

E. SOLUTION METHODOLOGY

The model is formulated as a linear integer program, and is solved using Lagrangian relaxation. The solution on a main frame computer averages under ten seconds.

---

[22] Determining the projected reenlistment rate at each bonus level is the purpose of this study.

F. MODEL USE

Because of the short run time, and the ease of input and interpretation of results, this model is extremely valuable to an analyst who must compare numerous alternative solutions, and perform sensitivity analysis of input variables. Although not specifically designed for use by budget analyst, the model can also be useful in budget development.

APPENDIX B. CALCULATION OF REENLISTMENT RATES

A. GENERAL

The purpose of the appendix is to explain how this study deals with four potentially complicating issues in the calculation of reenlistment rates. These issues are:

- MOS Migration
- Extensions
- Reenlistment Eligibility
- Early Reenlistments

How the study addresses these four issues has a profound impact on the calculation of the reenlistment rate. Therefore we start simply by defining how to calculate a reenlistment rate.

$$\textit{Reenlistment Rate } MOS_i = \frac{\textit{Number Soldiers Reenlisting in } MOS_i}{\textit{Number } \text{of } \textit{Soldiers Eligible}} \tag{7}$$

Each of the complicating factors potentially impacts on this rate calculation. The simplifying assumptions to prevent this are presented here.

B. MOS MIGRATION

MOS migration is when soldiers in an overstrength MOS reenlists into another understrength MOS. MOS migration is encouraged at the reenlistment point as a way to align the Army's force structure. The issue is how to count migrating soldiers in the calculation of reenlistment rates.

MOS migration effects the numerator of the reenlistment equation. There are four different ways to count migrating soldiers.

- Count in the numerator only soldiers in MOS_i who reenlist in MOS_i.
- Count in the numerator only soldiers from MOS_i who reenlist in MOS_i and those from all other MOS_j $i \neq j$ who reenlist for MOS_i
- Make the reenlistment decision a multinomial choice, to either reenlist for MOS_i, reenlist for any MOS_j $i \neq j$ or not reenlist.
- Count in the numerator soldiers in MOS_i who reenlist in any MOS_j including j.

By process of elimination, the study chooses the first method of calculation. The second method is rejected because there is no practical way to predict how many soldiers of other MOS's will choose to reenlist in MOS_i. The third choice, the multinomial choice, is rejected due to a technical aspect of the multinomial logit model. This solution technique works well only in cases in which there are three distinct choices. Here, two of the choices (to reenlist in MOS_i and to reenlist in MOS_j) are so similar as to render the technique ineffective [Ref. 35: p. 362]. The fourth option is rejected because it does not reflect the number of soldiers who remain in a MOS, which is vital information for the military reenlistment bonus model. Therefore the first option is selected. The benefit is this option keeps the model simple, and although there is some potential to underestimate the actual numbers of soldiers reenlisting for MOS_i it is the best option.

C. EXTENSIONS

Some researchers, such as Goldberg and Warner, treat extensions as a separate decision. They use a multinomial model of three choices (extend, reenlist, and leave the service) [Ref. 36: p. 17]. This study rejects this approach, and instead chooses to treat extensions as a deferred reenlistment decision. Therefore, only a soldier's final reenlistment decision is counts in the reenlistment rate calculation. This will case bias in the rate calculation only if soldiers extend in great numbers and for long periods. However, less than one in seven soldiers extend, and their primary reason for extending is to become reenlistment eligible. This method of treating extension is supported by the research by Cymrot. His conclusion is that the effects of extensions are small, (less than 1%) and he recommends that the inputs to the reenlistment models do not have to be modified to account for extensions [Ref. 37: pp. 44-46]. Therefore, extensions are ignored, at only a small cost to the accuracy of the model, and at a large benefit to the model simplicity.

D. REENLISTMENT ELIGIBILITY

This study counts all soldiers who reach their end of term of service (ETS) as eligible to reenlist. This is not the normal interpretation, as many soldiers are declared ineligible to reenlist as they do not meet the Army's minimum reenlistment standards. However, the difficulty with this approach is the data in the gain lose file designating reenlistment eligible soldiers is widely regarded as unreliable [Ref. 5: p. 26]. Any reenlistment rate based on this data is also unreliable.

Therefore the best approach is to declare all soldiers who reach ETS as eligible to reenlist. Since reenlistment eligibility standard have remained relatively unchanged over the past ten years, this is not an unreasonable approach. The estimation of the number of soldiers ineligible to reenlist than becomes a transparent part of the reenlistment rate computation.

E. EARLY REENLISTMENTS

Currently, soldiers are permitted to reenlist up to eight months prior to their ETS date.[23] This issue complicates the reenlistment rate calculation by changing the numerator of the reenlistment equation.

In his study Cymrot shows that there is no simple way to account for early reenlistments effect on the reenlistment rate, and that the forecast error of reenlistment rates is about 2% due to it [Ref. 38: p. 26]. This study recommends that soldiers are only counted as eligible to reenlist on one date, arbitrary set at six months prior to their ETS data.[24] This again greatly simplifies the model, although it cause the potential for some bias in the estimation. The bias is in the case of rising bonus levels, when soldiers who have previously decided not to reenlist change their minds due to a new, higher bonus level. In the case of falling bonus levels, there is no bias.

---

[23] Through FY87, first term soldiers were allowed to reenlist six months prior to the end of their service term, and all other soldiers were permitted to reenlist three months prior. Since FY 88, all soldiers are permitted to reenlist eight months prior to the end of their service term.

[24] 50% of soldiers reenlist eight to six months prior to their ETS, and 35% of soldiers reenlist six to three months prior, that the six month date is not unrealistic.

APPENDIX C. VARIABLES TO MEASURE INITIAL MOTIVATION FOR MILITARY SERVICE

The purpose of this appendix is to more fully explain a soldiers initial motivation for military service. This is part of the conceptual framework of the military decision-making process introduced in Chapter III.

The data for these variables comes from the Army gain loss file, except for the unemployment rate information which is from the Bureau of Labor Statistics.

ACF	Army College Fund (ACF) In a very interesting study of the Navy enlisted force, one researcher finds that educational programs reward military personnel leaving the service by providing what is in effect a negative reenlistment bonus, in the form of educational benefits that can only be used by a full time civilian student [Ref. 39: p. 2]. It is hypothesized here that a soldier motivated for military service by college money is less likely to reenlist after the first term.
ENLISTMENT BONUS	Studies show that soldiers receiving a *reenlistment* bonus at their first reenlistment point are less likely to reenlist once they reach their second reenlistment point [Ref. 40: p. 701]. Is there a similar effect for soldier receiving *enlistment* bonuses? If enlistment bonuses bring people into the service who otherwise do not enlist, then these soldiers may show a lower propensity to reenlist then other soldiers. The Army also uses enlistment bonuses to induce people to enlist in less popular job skills. These soldiers may be more likely to migrate to a new job skill at the end of their enlistment term.
ENLISTMENT TERM	One theory is that a longer enlistment term may indicate a stronger initial career intent on the part of the soldier. This is mitigated, however, because a soldier must enlist for four years to earn an enlistment bonus, and soldiers receiving enlistment bonuses may have less career intent.
PROGRAM	Enlistment Program Enlistment Program. This variable shows which enlistment or training program the soldier reenlists for. The purpose is to determine whether a soldier is job, training or education orientated. Studies show that soldiers in these different groups have different propensities to reenlist and also response differently to outside factors such as the state of the national economy [Ref. 40: p. 701]. The

enlistment program and training the soldier selects gives insight into the soldiers initial orientation.

AGE AT ENLISTMENT	Is there a correlation between age at enlistment, and enlistment motivation? One study by the RAND corporation shows a strong correlation between age at enlistment and first term attrition[25] [Ref. 41: p. vii]. It is hypothesized here that age at enlistment is also a predictor of enlistment intent.
AGE AT SEPARATION	Because soldiers enlist for different terms, age at separation is not exactly correlated to age at enlistment. Older soldiers are expected to reenlist at higher rates then younger ones.
EDUCATION	Education at enlistment. Initially, only a variable for education at reenlistment was included in this study (see Appendix D for discussion of the variable *Education*). However, education at enlistment can potentially explain a soldiers motivation for entering the service. Therefore, it is included here also.
DEPENDENTS	Dependents at enlistment. Similar to education, a soldiers dependent status at enlistment is included as a variable in this study.
PRIOR SERVICE	Has the soldier with prior military service followed by a break in service explored both the civilian and military opportunities available, and now indicated with his or her choice a strong career intention?
RESERVE TIME	Likewise, is a soldier who is serving in the Reserves or National Guard and then decides to come on active duty more career oriented then the average soldier?
YOUTH PROGRAM	Participation in military youth programs such as high school ROTC may indicate that this individual, like reserve and prior service soldiers, has made comparisons of both civilian and military options available from a perspective not available to the average person.
HOMETOWN	Location, along with the economic conditions at that location are strongly related to enlistment propensity according to one study [Ref. 42: p. 230]. Hometown information is converted to regional information for use in this variable. The regions are further combined, so that five large regions are formed. States in each region have soldiers with similar reenlistment rates.
UNEMPLOYMENT RATE	The unemployment rate is examined as an indicator of an individuals motivation to enter the military. Two different unemployment rates are used here. One is the

---

[25] Soldiers under the age of 18 show significantly higher first term attrition rates then older soldiers.

average state unemployment rate for the 13 months prior to the soldier enlisting. The other is the national rate for the same period. The justification for using these rates comes from a study on the sensitivity of first term Navy reenlistments to changes in unemployment and relative wages [Ref. 40: p. 698]. Unemployment data comes from the Bureau of Labor Statistics [Ref. 43: p. 8].

APPENDIX D. VARIABLES TO MEASURE THE SOLDIERS SUCCESS IN THE SERVICE

The purpose of this appendix is to further describe variables which measure a soldiers success in the service, and his or her satisfaction with military life. This is part of the conceptual framework of the reenlistment decision-making process introduced in Chapter III. All data comes from the Army gain loss file except where noted.

CHARACTER OF SERVICE	At each reenlistment point, the soldier receives a character of service. This is a gross indicator of previous performance, because if the character of service is anything less than honorable, the soldier is not permitted to reenlist.
PROMOTION RATES	Promotion rates of soldiers compared to their peers within their military occupation specialities appears to be the best way to measure a soldiers success within the military. Soldier's enlisted evaluation report scores and skill qualification test scores also look promising, but data is not available. The use of promotion rates as an indicator of success in the military is well supported in studies such as a RAND study [Ref. 16: p. v]. The method of calculating promotion rates is the same used by Warner in his masters thesis [Ref. 17: p. 38].
AFQT SCORE	Armed Forces Qualification Test. Two studies, one by the RAND Corporation, and one by an NPS student use intelligence and education scores to predict promotion rates. AFQT, plus the following three variables (mental test category, GT score, and education level) are measures of intelligence and education, although each comes with serious and well documented shortcomings as a measurement tool. Additionally, the results of studies which use these variables as predictors are not particularly strong [Ref. 16: p. 3] [Ref. 17: p. 120]. Despite its shortcomings, the Army makes frequent use of this measure of intelligence.
MENTAL TEST CATEGORY	This variable is also one of those used to predict promotion rates. Mental test category is a discrete version of the AFQT, ranging from 1 (highest) to 5 (lowest). Each category is further broken into sub-categories. The mental test category is hampered by the same inconsistencies described for the AFQT.
GT TEST SCORE	General-Technical Test Score on the Armed Forces Vocational Aptitude Battery. Another of the variables

	used to predict promotion rates. The Army uses this test score data to measure trainability.
EDUCATION LEVEL	The final variable used to predict promotion rates. The problem with the measure of education level available in the data base is that it does not distinguish between soldiers who are high school graduates and those who earn a high school equivalency credential (GED).[26]
CHANGE IN EDUCATION	Since the study examines education at enlistment, and education at the reenlistment point, it also examines whether soldier who have improve their education level during their enlistment term have different enlistment probabilities then those who do not.
YEARS-OF-SERVICE	An Army Research Institute researcher discusses the use of tenure in the service as predictor of organizational commitment and reenlistment propensity [Ref. 44: pp. 5-6]. He measures tenure with four factors: years-of-service, status, rank and increasing responsibility. Data is available on years-of-service and rank.
CURRENT RANK	A second measure of tenure.
DUTY LOCATION	This study uses duty location as a quality of life variable. A study of first term reenlistment decisions finds that Army enlistees who are stationed overseas have a higher reenlistment rate, and those stationed in the northeast United States have a lower reenlistment rate then average [Ref. 8: p. 23]. The duty station is converted into regional or overseas location.
DEPENDENT STATUS	Researchers note that quality of life issues are relatively insignificant for the first term soldier [Ref. 20: pp. 11-14]. The reason may be that many first term soldiers do not yet have families, while later term soldiers do. Soldiers with families, or who support dependents should reenlist at higher rates then single soldiers do. This thesis defines a soldier as having dependents if he has any legal dependents, whether they are children, parents, or other relatives.
CHANGE IN DEPENDENTS	Does a soldiers who starts his or her family while in the military display different reenlistment propensity then single soldiers, or those who entered with families? This variable addresses the issue.

---

[26] Education level data which distinguishes between GED graduates and high school diploma graduates is only available from 1985 on.

APPENDIX E. VARIABLES TO MEASURE A SOLDIERS POTENTIAL IN THE CIVILIAN SECTOR

The purpose of this appendix is to more fully explain a soldiers evaluation of his or her potential in the civilian sector. This is part of the conceptual framework of the reenlistment decision making process introduced in Chapter III. The data is this group comes from the appropriate government agency, and from the Army gain loss file.

RACE	The study includes race and sex as surrogates variables to describe a soldier's evaluation of his or her potential in the civilian sector verses the military. Researchers find higher reenlistment rates among black soldiers than white soldiers. The researchers hypothesis this is due to several factors, such as insufficient job opportunities for blacks in the civilian sector as compared to military career options, and enhanced promotion opportunities in the military [Ref. 14: pp. 29-30]. Therefore race becomes an indicator of differing opportunities available to soldiers in civilian sector and the military.
ETHNIC GROUP	For similar reasons as for race, a soldiers ethnic group is included as a variable.
SEX	Studies also note higher reenlistment rates among women then men for first term soldiers[27] [Ref. 14: p. 29]. Again, researchers hypothesis this represents more opportunities for women in the military then they find in the civilian sector [Ref. 14: pp. 29-30].
JOB TYPE	The purpose of this variable is to attempt to capture different civilian opportunities for differing job categories. Most researchers agree that soldiers with "high tech" training have greater civilian opportunities than do other soldiers [Ref. 2: p. 8] [Ref. 4: p. 253]. This variable also captures the expected lower bonus response rates for jobs that are risky or dangerous [Ref. 4: p. 231]. The Army's administrative grouping of job skills into categories called career management fields (CMF), which we do not use because CMF's often group occupations with little in common [Ref. 5: p. 4].[28] This study uses instead modified groupings from

---

[27] Women have a higher attrition rate then men during the first term. However if they complete the first term, women reenlist at a higher rate then men.

[28] For example, CMF's group job skills as diverse as a cannon crewman and a Pershing missile electronics specialist into the same category.

	the Department of Defense *Occupation Conversion Manual* [Ref. 45: pp. 9-17].
UNEMPLOYMENT RATE	Numerous studies find unemployment rates positively correlated with retention rates, and that unemployment rates reflect civilian employment opportunities [Ref. 6: p. 16]. Additionally, the unemployment rate, (along with GNP and CPI) indicate the health of the national economy [Ref. 2: p. 54]. A study for the U. S. Navy titled "The Sensitivity of First Term Navy Reenlistment to Changes in Unemployment and Relative Wages" addresses the wide range of issues dealing with which unemployment rates to use[29] [Ref. 40: p. 54]. This study uses two, the state unemployment rate for the 13 months prior to the soldiers enlistment (discussed in Appendix C), and the national unemployment rate for the three quarters prior to the soldier making his reenlistment decision. Unemployment data comes from the Bureau of Labor Statistics [Ref. 43: p. 8].
C/M WAGE INDEX	Civilian Military Wage Index. Surprisingly, studies do not find civilian military pay indexes to be explanatory of the reenlistment decision making process. Only one Navy study finds them to be significant predictors of reenlistments [Ref. 36: p. 32]. Numerous others find this not to be true [Ref. 14: p. iii] [Ref. 40: p. 707] [Ref. 8: pp. 35-36] [Ref. 9: pp. 40-43]. The difficulty here is trying to measure the civilian earning potential of soldiers. One approach is to use veterans earnings as a way to estimate the earning potential of soldiers in the civilian sector. However this introduces selection bias into the data, because veterans who choose to leave the service do so because they expect higher civilian earnings than those who stay. Therefore any estimate of civilian wage potential based on veterans earnings is upwards biased [Ref. 11: p. 203] [Ref. 46: p. v]. Another difficulty with measuring civilian pay opportunities of soldiers is matching military skills with skills found in the civilian sector. Despite the above shortcomings, this study includes the civilian military wage index as a variable. The source of data is the Bureau of Labor Statistics [Ref. 43: pp. 115-177].
CPI	Consumer Price Index. Like unemployment and gross national product, CPI is a general measure of the state of the national economy, and therefore employment

---

[29] The issues break down into whether to use national, regional, or local unemployment rates; whether to use the rates for all workers or those for the 17-24 age group; and how much should the effects of unemployment be led or lagged.

	opportunity. The source of data is the Labor Statistics [Ref. 47: pp. 13-16].
GNP	Gross National Product. GNP also indicates the health of the national economy, and therefore indicates the civilian employment prospects of military personnel. None of the studies reviewed for this paper include GNP as a variable, although GNP is the most frequently used measure of the state of the national economy. GNP data is from U. S. Department of Commerce [Ref. 48: p. 3].
CIVILIAN JOB GROWTH	This study hypothesizes that the percentage growth in civilian jobs is a more accurate indicator of actual employment opportunities than is the unemployment rate. Data come from the Bureau of Labor Statistics [Ref. 43: p. 30].

APPENDIX F. REENLISTMENT POLICY VARIABLES

The purpose of this appendix is to more fully explain the reenlistment policy variables in this study. The variables are part of the conceptual framework of the reenlistment decision making process of Chapter III. Data in this section comes from the Army gain loss file except where noted.

RETIREMENT SYSTEM	The purpose of this variable is to account for changes in the retirement system made four years ago. Soldiers enlisting before this date received benefits under the old retirement system. The new retirement system is less generous then the old one [Ref. 14: pp. 29-30].
YEARS TO RETIREMENT	One of the strongest predictors of reenlistment behavior is the number of years to retirement. However, this variable is most useful in predicting Zone B and Zone C reenlistment rates. The years to retirement have little influence on Zone A soldiers, with the major impact not felt until the seventh year [Ref. 14: p. 17].
RMC	Real Military Compensation. RMC is a measure of compensation that accounts for the fact that not all of a soldiers income is in the form of direct pay. RMC accounts for the housing and substance allowances that soldiers receive either in cash or in kind (in the form of government housing). RMC also counts as income the tax advantage a soldier gets because housing and substance payments are not taxable. Due to the fact that the military compensation system is sufficiently complex, there is considerable evidence that soldiers systematically and significantly undervalue their compensation [Ref. 41: p. vi]. Changes in pay rates, rather than actual pay rates where used in this study.
ADJUSTED RMC	This variable takes into account how pay (and other forms of military compensation) keep pace with inflation.
BONUS PAYMENTS	The bonus payment level is the policy variable Army policy makers can most easily manipulate. Since bonuses are paid to soldiers in job skills with low retention rates, normally the presence of a bonus indicates that the job skill is in high civilian demand or is an unpopular or demanding job. Bonus payment data comes from the Force Alignment branch of the U. S. Army Total Army Personnel Command.

TYPE BONUS PAYMENT	The method of computing the amount of a reenlistment cash bonus has not changed since 1974. However, the method of payment has changed. From April 1979 to January 1982, the cash bonus was paid to the soldier in a lump sum on the day of reenlistment. However, in 1982 the method changed from a lump sum to a one-half lump sum payment, with the remainder of the bonus paid in yearly installments. Studies show that the full lump sum payment induces more soldiers to reenlist then the alternate payment system [Ref. 6: p. 6] The data base includes records of soldiers under both payment systems. Bonus type data comes from the Force Alignment branch of the U. S. Army Total Army Personnel Command.
SKILL MIGRATION	The Army permits selected soldiers to change job skills at the reenlistment point. The force alignment needs of the dictate the number of soldiers who change job skills. The Army offers soldiers in overstrength MOS's the opportunity to change to understrength MOS's. These soldiers normally do not receive a bonus, however their reward for changing MOS's is increased promotion opportunity in the new MOS. This variable indicates whether the soldier is in an overstrength MOS and eligible to reenlist. Migration opportunity data comes from the Force Alignment branch of the U. S. Army Total Army Personnel Command.
PROMOTION FORECAST	An earlier variable looks at the promotion rate of a soldier respect to his peers. This variable looks at the promotion rate as a force alignment variable which the Army manipulates. Promotion forecasts come from the Force Alignment branch of the U. S. Army Total Army Personnel Command.
ELIGIBILITY	Reenlistment eligibility criteria change over time. The data base contains a variable coding reenlistment eligibility, however this designation is highly suspect [Ref. 5: p. 26]. We are not able to independently determine from the data records whether a soldier is eligible to reenlist, as reenlistment eligibility depends partially on discipline and performance records not available for this study. Therefore, this variable measures which set of reenlistment eligibility criteria is in effect at the time the soldier reenlists.
REENLISTMENT SYSTEM	The purpose of this variable is to attempt to quantify how liberal the reenlistment system is in giving a soldier his or her reenlistment choice of training or duty assignment. This study subjectively assigned values to this variable, based on interviews with the reenlistment managers at the U. S. Total Army Personnel Command. The general feeling is that from FY82 through

FY83, the reenlistment system was moderately responsive to soldier's needs. From FY84 through FY87, the reenlistment system was less responsive to soldier's needs, and during FY88 and FY89 it has been more highly responsive to soldier's needs. This assessment is due to changes in the reenlistment system that occurred on 1 October 1983, and in 1 April 1988.

APPENDIX G. MISSING DATA

A. PURPOSE

The purpose of this appendix is to show the amount of missing data present in the data set after cleaning, and to demonstrate why no further cleaning of the data set is required.

B. MISSING DATA AFTER CLEANING

Table 9 contains a listing of the 30 categorical variable, and the amount of missing data present after cleaning. The amount of remaining missing data ranges from 0-7.8%, with 23 variables missing less than 1%.

C. RANDOM MISSING DATA

To determine if further cleaning of the data is necessary, the data set is examined to see if the observations with missing data are a random sample of the data set. If they are, then eliminating the observations with missing data will not change the results of the analysis, and additional cleaning will not be needed.

First, the number of observations with at least one missing value is calculated, using the ten variables from Table 9 with the most missing data. The results are in Figure 16.

```
  DATA                                 CUMULATIVE  CUMULATIVE
 MISSING    FREQUENCY     PERCENT      FREQUENCY     PERCENT
-------------------------------------------------------------

   NO          69570        91.8          69570        91.8
   YES          6208         8.2          75778       100.0
```

Figure 16. Number of Observations With Missing Values

As can be seen, only 8.2% of all observations have one or more missing values. This amount is acceptable, provided the observations with missing values are a randomly distributed throughout the data set. To determine this, we test the hypothesis that the reenlistment rate for the those with missing data is the same as the reenlistment rate for those without missing data. Figure 17 gives the reenlistment rates.

Table 9. MISSING DATA FOR CATEGORICAL VARIABLES

Variable Name	Percentage of Data Missing
AFC	6.56%
Enlistment Bonus	0.00%
Enlistment Term	0.00%
Enlistment Program	7.88%
Age at Enlistment	0.02%
Age at Separation	0.01%
Prior Service	5.12%
Reserve Time	0.00%
Youth Program	0.00%
Hometown	0.00%
Education at Enlistment	0.04%
Education at Reenlistment	0.01%
Change in Education	0.04%
Dependent Status at Enlistment	5.75%
Dependents at Reenlistment	0.03%
Change in Dependents	5.76%
Character of Service	0.52%
Mental Test Category	1.24%
Years of Service	0.07%
Current Rank	0.00%
Duty Location	0.53%
Race	0.03%
Ethnic Group	0.01%
Sex	0.00%
Job Type	0.02%
Retirement System	0.00%
Number of Years to Military Retirement	0.00%
Type of Bonus Payment	0.00%
Job Skill Migration	6.49%
Reenlistment Bonus	0.00%

```
                 RESPONSE PROBABILITIES


       DATA            RESPONSE NUMBER
      MISSING     NO REENLIST     REENLIST
      ------------------------------------
        NO          0.617824      0.382176
        YES         0.619845      0.380155
```

Figure 17. **Reenlistment Rates for Observations With Missing Data**

Obviously, the reenlistment rate for those observations missing data is very close to that for those not missing data. To show this formally, we test the hypothesis:

$$H_0\colon P_1 = P_2 \tag{8}$$

$$H_1\colon P_1 \neq P_2 \tag{9}$$

Where P_1 is the probability of reenlisting of an observations without missing data, and P_2 is the probability of reenlisting of an observations with missing data. The test statistic is:

$$T = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{n_1 n_2 C_1 C_2} \tag{10}$$

where N is the total number of observations, n_1, n_2, C_1, C_2 are the row and column totals and O_{11}, O_{22}, O_{12}, O_{21} are the cell frequencies.

The critical region is to reject H_0 at $\alpha = 0.05$ if T exceeds $X^2_{1-\alpha}$, the $(1 - \alpha)$ quantile of a chi-square random variable with 1 degree of freedom [Ref. 27: pp. 145-146]. Since $T = 0.09866$ is much less than $X^2_{1-.05} = 7.879$, we do not reject the null hypothesis. The level of significance of the test is greater then $\hat{\alpha} = 0.25$

Therefore, since the missing values appear to be randomly distributed throughout the data set, further cleaning of the data is not required.

APPENDIX H. LOG-LINEAR MODELS

The purpose of this appendix is to explain the use of log-linear models in the study of categorical data sets. The log-linear model is analogous to the familiar analysis of variance (ANOVA) techniques, except that log-linear models are for dichotomous response variables, where the ANOVA is for continuous response variables. Both are for use with categorical explanatory variables.

The standard log-linear model is Equation 11, where p_i, p_j, p_k are the probabilities associated with the different variables.

$$Rate = Ap_ip_jp_k \tag{11}$$

Taking the natural logarithm of this equation yields Equation 12.

$$Rate = \ln A + \ln p_i + \ln p_j + \ln p_k \tag{12}$$

The SAS statistical procedure CATMOD uses a maximum likelihood estimate solved by a iterative proportional fitting procedure to yield estimators that are the best asymptotic normal estimators [Ref. 49: p. 35]. The properties of iterative method of proportional fitting of the log-linear model are summarized from Bishop [Ref. 26: p. 83].

- It always converges to the required MLE.
- A stopping rule is available to ensure the desired accuracy is obtained.
- Starting values may be set for the estimates.

The SAS categorical modeling procedure performs hypothesis tests to determine if the estimated parameters are significantly different from zero. The test statistic is a Wald statistic, which is approximated by a chi square distribution [Ref. 49: p. 35].

APPENDIX I. LOGISTIC REGRESSION

The purpose of this appendix is to describe the regression techniques used in this thesis.

The key issue in selecting the regression techniques is the dichotomous response variable. Soldiers make only one of two mutually exclusive reenlistment decisions, ether to reenlist or leave the service. [30]

Since the response variable is binary, the desired result of the regression equation is the probability of success (reenlistment) of a given soldier.

$$P_i = P(Y_i = 1) \qquad (13)$$

Where $Y_i = \{0, 1\}$.

To apply a ordinary least squares regression to this, the following interpretation is made. The general form of the linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (14)$$

If P_i is the probability that $Y_i = 1$, then:

$$E[Y_i] = P_i = \beta_0 + \beta_1 X_i \qquad (15)$$

if $E[\varepsilon_i] = 0$. This is the linear probability model [Ref. 50: p. 12] [Ref. 35: p. 756].

There are a number of reasons why using ordinary least squares regression is not adequate for models having categorical response variables.

- By definition, the probability P_i in Equation 13 must take on values between 0 and 1. However, using the linear regression model, the P_i can fall outside the 0, 1 range. Figure 18 shows this where the solid line represents an actual probability function, and the dashed line represents a linear approximation to it. In this example, the linear approximation goes outside the 0, 1 range for admissible $\beta_0 + \beta_1 X_i$ [Ref. 51: p. 4].

---

[30] Some researchers study a multinomial reenlistment choice, however for reasons described in Appendix B, this study uses a dichotomous response variable.

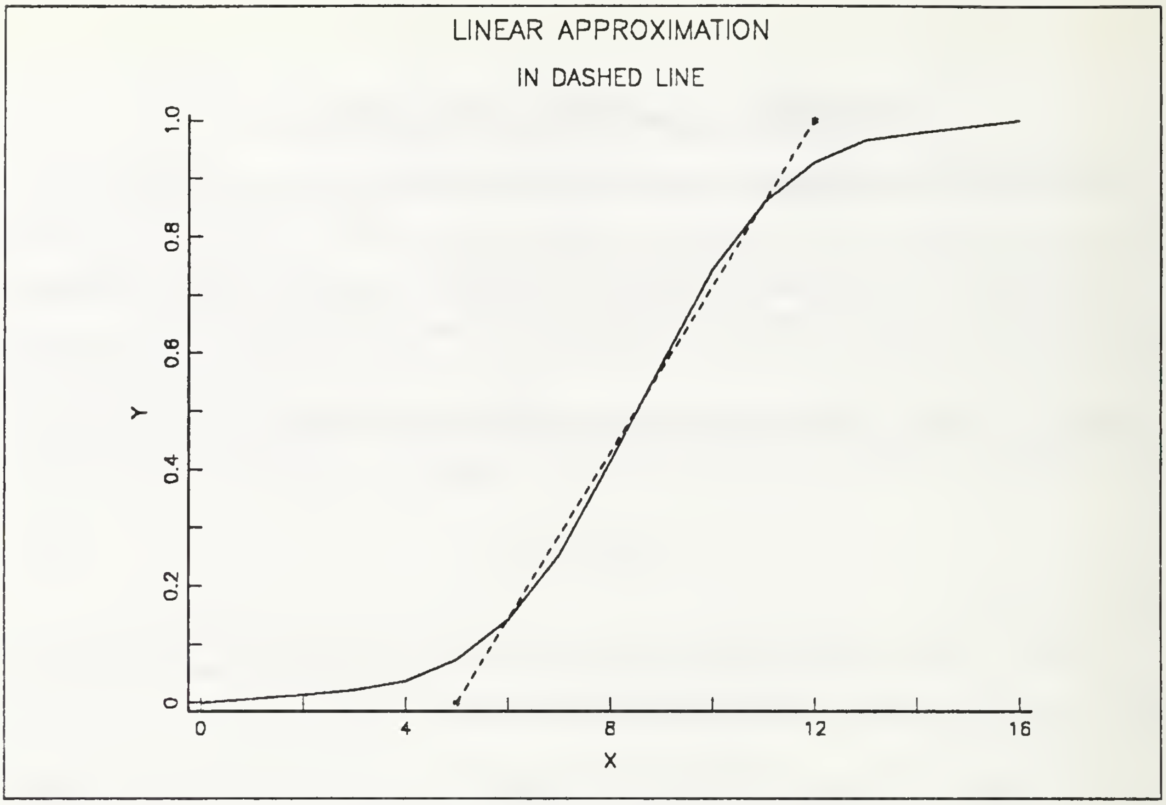

Figure 18. Linear Approximation to a Probability Function

- Linear regression uses the assumption of constant variance of errors, $E[\varepsilon_i^2] = \sigma^2$. However, the variance of the error term for a binary variable, where each observation is assumed to be a Bernoulli trial, with probability of success P_i is:

  $$Var[\varepsilon_i] = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_i X_i) \tag{16}$$

  Since the variance of the errors depends on the observation, the ε_i do not have constant variance. Use of ordinary least square regression models produces inefficient estimates and imprecise predictions [Ref. 35: pp. 419-422].

- The assumption that the Y_i are normally distributed is not valid with binary data. This is obvious, as the Y_i are either 0 or 1. Since they are not normally distributed, no estimation that is linear in Y_i is efficient [Ref. 35: pp. 419-422].

- The usual tests of significance for the estimated coefficient do not apply when using ordinary least squares on observations with binary response variables; estimated standard errors are not constant, and R^2 does not have its usual interpretation [Ref. 35: pp. 419-422].

The solution to the above problems are transformations. The two most widely used transformation are the probit and the logit transformations. The probit transformation, which is based on the normal CDF is:

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{L_i} e^{-t^2/2} dt \qquad (17)$$

The logit transformation, which is used because of its close approximation to the normal CDF is:

$$P_i = \frac{1}{1 + e^{-L_i}} \qquad (18)$$

Both of these transformations work well when there are sufficient repeated observations available (when the explanatory variables are categorical). If, however, there are few repeated observations (continuous explanatory variables) then a maximum likelihood estimation of the logit model is used.[31] The data for the model is shown in Figure 19.

DATA

NUMBER OF TRIALS IN OBSERVATION I	NUMBER OF SUCCESSES IN OBSERVATION I	EXPLANATORY VARIABLES
M	S	X1 X2 ... XN
M	S	X1 X2 ... XN
.	.	. .
.	.	. .
.	.	. .
M	S	X1 X2 ... XN

Figure 19. Data Format for Logistic Regression

In this case the explanatory variables are continuous, and there is only one trial per observation ($M_i = 1$) and S_i is either 1 or 0 (success of failure). [Ref. 35: pp. 419-422]

---

[31] While the logit transformation is somewhat arbitrary, it is selected because it is simple, tractable and well behaved even when the normality of L_i is violated.

The following discussion of the development of logistic regression is summarized from Judge [Ref. 35: pp. 425-436] and Nerlove [Ref. 51: pp. 14-22]. Using the binomial distribution, the probability of a success in observation i is defined as:

$$P\{N_i = s\} = \binom{M_i}{s} P_i^s (1 - P_i)^{M_i - s} \tag{19}$$

where $M_i = 1$ and $s = 1$

The logit transformation is:

$$P_i = \frac{1}{1 + e^{\underline{X}_i \underline{\beta}}} \tag{20}$$

where:

$$\underline{X}_i \underline{\beta} = \sum_i X_{ij} \beta_j \tag{21}$$

The maximum likelihood function is:

$$L = \prod_i \binom{M_i}{S_i} P_i^s (1 - P_i)^{M_i - S_i} \tag{22}$$

Following the procedures for computing a maximum likelihood estimator in Larsen [Ref. 52: p. 262]. First take the natural log of the likelihood function, and substitute the expressions for P_i and $1 - P_i$.

$$ln\ L = \sum_{i=1}^{k} ln \binom{M_i}{S_i} - S_i\ ln\,(1 + e^{\underline{X}_i\underline{\beta}}) + (M_i - S_i)\,[\underline{X}_i\underline{\beta} - ln(1 + e^{\underline{X}_i\underline{\beta}})] \tag{23}$$

The next step is to take the derivative and set equal to zero, however this is not possible as the derivative is non-linear in the estimators. Instead, a Newton-Raphson method is used to find a numeric solution to the problem using an iterative procedure. The initial conditions are:

$$\underline{X}_i\underline{\beta}^0 = ln\left[\frac{M_i - n_i + \frac{1}{2}}{n_i + \frac{1}{2}}\right] \tag{24}$$

The first step of the iteration is to compute the weights:

$$w_i = \left[\frac{M_i e^{\underline{X}_i\underline{\beta}^0}}{\left(1 + e^{\underline{X}_i\underline{\beta}^0}\right)^2}\right]^{\frac{1}{2}} \tag{25}$$

$$u_{ij} = w_i x_{ij} \tag{26}$$

$$u_i = \frac{M_i - n_i}{w_i} - \sqrt{M_i e^{\underline{X}_i\underline{\beta}^0}} + \sum_{j=1}^{p} u_{ij}\beta_j^0 \tag{27}$$

The nest step is to perform a least square regression of dependent variables Y_i and the weighted dependent variables $U_{i1}, \ldots, U_{ip}$

$$\underline{\beta} = (U^T U)^{-1} U^T Y \tag{28}$$

Next, the estimates $\underline{\beta}^0$ are updated.

$$\underline{\beta}^0 = \underline{\beta} \tag{29}$$

$$\underline{x}_i\underline{\beta}^0 = \sum_{j=1}^{p} x_{ij}\beta_j \tag{30}$$

The procedure is continued until the estimates converge.

Using this procedure, the probability of success with a given set of explanatory variables is:

$$P_i = \left[\frac{1}{1 + e^{\underline{X}_i\underline{\beta}}}\right] \tag{31}$$

The above discussion is summarized from Judge [Ref. 35: pp. 425-436] and Nerlove [Ref. 51: pp. 14-22].

The statistical package of this study is the LOGIST procedure of the SAS statistical package [Ref. 53: pp. 181-202]. The procedure uses the maximum-likelihood estimates described above. Some specifics on the assumptions of the procedures, and the test statistics are:

- The assumption of the binary model is that the probability that $Y_i = 1$ is given by Equation 31.
- The response variable can be nominally scaled.
- The Logit model has few assumptions, and is robust to the assumptions of ordinary least squares regression.
- The logit transformation can be applied to a multivariate setting. This is justified, because the marginal distributions of the multivariate logit transformations are themselves logit transformations.
- The SAS LOGIST procedure examines two way interactions between variables, but higher order interactions are assumed to be zero.
- The form of the residuals is undetermined, however the transformed residuals should be approximately normally distributed.
- Test of hypotheses and confidence intervals in the SAS LOGIST procedure are constructed from estimates of the asymptotic covariance matrix using Wald statistics. These rely on the asymptotic nature of the maximum likelihood estimator. The confidence intervals could also be determined using a bootstrapping (resampling) procedure developed by Efron. [Ref. 54: pp. 5-18].
- The R statistic is similar to the multiple correlation coefficient in the normal setting after a correction is made to penalize for the number of estimated parameters.
- The SAS LOGIST procedure has a forward stepwise regression option, which is used in this study. Where a least squares stepwise regression uses a f statistic for variable selection, the SAS LOGIST procedure uses a Rao's efficiency score statistic. Similar to least squares regression, care must be taken in using the stepwise SAS LOGIST procedure. If arbitrarily applied without proper safeguards, a stepwise procedure can lead to an inaccurate model. One of the most effective methods to ensure performance of a stepwise procedure is to cross-validate the model. These issues are discussed in more depth in Freedman. [Ref. 55: p. 152].
- If a variable is a linear combination of other variables already in the model, then it will not be added to the model in the stepwise SAS LOGIST procedures.
- Finally, a SAS LOGIST NOFIT procedure is used as a diagnostic tool prior the fitting of models using stepwise procedures. This procedure tests the null hypothesis that all regression coefficients are zero. The NOFIT option is useful in finding out if any modeling is worth while at all.

The above are summarized from Judge [Ref. 35: pp. 425-436], Nerlove [Ref. 51: pp. 14-22], and Harrell [Ref. 53: pp. 181-202].

APPENDIX J. CLUSTER ANALYSIS RESULTS

This appendix gives the results of the clustering of cells, which is described in Chapter V. The soldier population is first partitioned into 1080 cells, and then in a two step procedure this number is reduced to thirty-six cells. The assumption is that each of these cells is a grouping of soldiers with a similar probability of reenlisting. The assumption is tested in this appendix, using a non-parametric goodness-of-fit test.

The cells are coded to identify which groups of soldiers belong to them. The coding is by the seven variables used to define the cells. Those variables (in the order in which they appear in the coding) are as follows:

- Term of Enlistment
- Sex
- Rank
- Dependents
- Race
- Region
- Job Type

The number in each position of the coding represents the category of the variable represented. The possible categories for each variable are:

• Term of Enlistment	(2-two years, 3-three or more years)
• Sex	(1-male, 2-female)
• Rank	(3-E3 or below, 4-E4, 5-E5 and above)
• Dependents	(1-no dependents, 2-married or single with dependents)
• Race	(1-white, 2-black, 3-other)
• Region	(1-northeast, 2-mid-atlantic, 5-south, 7-midwest, 8-west)
• Job Type	(1-low, 2-medium, 3-high civilian opportunity)

An asterisk in the coding means that the given all categories in the given variable are combined, plus all categories of all remaining variables in the hierarchical structure are combined. Two numbers with parentheses around them represent two categories grouped together.

Three examples of this coding scheme are provided. The first, in Equation 32, represents all soldiers who enlisted for three or more years, are male, are of rank E4, with dependents, are of a ethnic group of other than white or black, are from the south, and are in an MOS that provides a medium level of civilian opportunity.

$$3\ 1\ 4\ 2\ 3\ 5\ 2 \tag{32}$$

The coding of Equation 33 represents all soldiers who enlisted for two years and are female. (The asterisk means that the cell contains soldiers in all categories of the variables **RANK, DEPENDENTS, RACE, REGION** and **JOB TYPE**.)

$$2\ 2\ * \tag{33}$$

The coding of Equation 34 represents all soldiers who enlisted for two years, are male, are of rank E3, and are either black or in the other ethnic code classification.

$$2\ 1\ 3\ 1\ (2\ 3)\ * \tag{34}$$

Tables 10 and 11 give the composition of each cell.

Figures 20 and 21 give the expected reenlistment rate for each of the 36 cells, and the number of observations of a sample of 75,778 total.

We now test the assumption that a cell is a grouping of soldiers with a similar probability of reenlistment. To do this, we use the validation data we have been saving. A chi-square goodness-of-fit test is preformed, testing the assumed distribution function on each cell of the validation data. The hypothesis is that the observations in a given cell are distributed $Binomial(n, p)$ where p is the estimated reenlistment rate given in Figures 20 and 21. In the test statistic in Equation 35, O_1 is the observed number of soldiers reenlisting, O_2 is the observed number of soldiers leaving the service, E_1 is the expected number of soldiers reenlisting, and E_2 is the expected number of soldiers leaving.

$$T = \sum \frac{(O_j - E_j)^2}{E_j} \tag{35}$$

The decision rule is to reject H_0 if T is greater than $X_{1-\alpha}$, the $(1-\alpha)$ quantile of a chi-square random variable with 1 degree of freedom. In this test, $X_{1-\alpha} = 3.841$ for $\alpha = 0.05$ and $X_{1-\alpha} = 10.83$ for $\alpha = 0.001$. Figures 20 and 21 list the reenlistment rate for the each validation cell and the *T* statistic for each cell. For any goodness-of-fit test, the null hypothesis is rejected if the sample size is allowed to get large enough [Ref. 27: pp. 190-191]. Cells 15 and 55 show this, as they are cells with larger sample sizes, and moderate differences in probability (less than one percent), yet they have large *T* statistics. Therefore, even though some of the tests reject the null hypothesis, the overall effect of the chi-square test is to confirm the distributional assumptions of the cells. Therefore, we conclude that we have partitioned the population into cells of soldiers with similar reenlistment probabilities.

Table 10. CLUSTER RESULTS BY ZONE

CELL #					
Cell 1	22*	315111*			
Cell 2	2132*	2131(23)*	32411(18)*		
Cell 3	21311*	3131*			
Cell 5	2142*	32411(27)*	32421*		
Cell 6	21411*	323*	3132*		
Cell 7	2141(23)*				
Cell 8	215*	3241(23)*			
Cell 12	324115*	3151122	3152113		
Cell 15	3242(23)*				
Cell 16	32511*				
Cell 17	3251(23)*	3152152	3152172	3142252	
Cell 18	3252*				
Cell 22	315112(13)	3151183			
Cell 24	315115*	315117*			
Cell 26	315118(12)	314132(13)	314137(13)	3141151	3141153
Cell 26 (cont)	3141171	3141172			
Cell 28	3151(23)*	315211(12)	3152122	315215(13)	315217(13)

Table 11. CLUSTER RESULTS BY ZONE (CONTINUED)

CELL #					
Cell 31	315212(13)				
Cell 37	315218*	314235(13)	314221*		
Cell 38	3152(23)*				
Cell 39	314231*	314232(23)			
Cell 41	3142321	314237(13)			
Cell 43	3142352	3142372			
Cell 46	314238(12)	314212*	314215(23)	314218*	314122(13)
Cell 46 (cont)	314127(13)				
Cell 47	3142383	3142113			
Cell 49	3142221	3142151			
Cell 50	3142222				
Cell 51	3142223	314227(13)	3142172	3141252	
Cell 52	3142251	314228*			
Cell 54	3142253	3142272			
Cell 58	314211(12)	3141322	314135(12)	314128*	
Cell 63	314217(13)				
Cell 66	314131*	314111*			
Cell 70	3141353	3141122	3141152	3141173	
Cell 72	3141372	314121*			
Cell 73	314138*	3141121	3141123	314118*	
Cell 76	3141222	314125(13)	3141272		

	MODEL BUILDING DATA		VALIDATION DATA		
CELL	SAMPLE SIZE	PERCENT REENLISTING	SAMPLE SIZE	PERCENT REENLISTING	T
1	1013	.311945	970	.312371	0.00
2	928	.246767	950	.251579	0.11
3	5409	.081161	5439	.080530	0.02
5	3094	.347447	3274	.367440	6.04
6	4583	.130700	4406	.128007	0.35
7	458	.283843	405	.244444	3.12
8	1575	.532698	1582	.530973	0.03
12	481	.484407	467	.456103	1.46
15	1845	.595122	1834	.585605	183.
16	407	.449631	380	.410526	2.39
17	834	.701439	791	.701643	0.00
18	880	.643182	886	.638826	0.07
22	1759	.363275	1834	.371865	0.62
24	1190	.398319	1138	.384007	0.93
26	4260	.276291	4290	.286713	2.46
28	3303	.635966	3042	.638067	0.06
31	1714	.578763	1684	.592043	1.18
37	910	.550549	928	.607759	12.5
38	1786	.800112	1809	.799889	0.00
39	244	.606557	245	.526531	6.65
41	368	.472826	421	.441805	1.64
43	232	.607759	234	.585470	0.50
46	10266	.433275	10374	.427607	1.23
47	470	.340426	469	.432836	18.0
49	1331	.514651	1433	.501047	1.12
50	443	.668172	432	.618056	4.86

Figure 20. Number of Observations and Reenlistment Rates by Cell

	MODEL BUILDING DATA		VALIDATION DATA		
CELL	SAMPLE SIZE	PERCENT REENLISTING	SAMPLE SIZE	PERCENT REENLISTING	T
51	2407	.560033	2310	.553247	341.
52	930	.600000	923	.582882	1.13
54	743	.647376	802	.665835	1.25
58	1452	.404270	1449	.402346	0.02
63	1604	.459476	1559	.463117	0.11
66	2324	.206540	2269	.228735	6.53
70	2635	.310816	2701	.276564	14.9
72	259	.374517	287	.324042	3.18
73	10120	.246739	10029	.246086	0.05
76	3621	.483568	3610	.497230	2.53

Figure 21. Number of Observations and Reenlistment Rates by Cell (Continued)

APPENDIX K. REGRESSION ANALYSIS RESULTS

The purpose of this appendix is to present the regression analysis results for each cell. A stepwise logistic regression procedure estimates the coefficients. A description of the method of inclusion of variables appears in Appendix I. Except for the intercept terms, all coefficients are significant at the $\alpha = 0.05$ level. Those intercepts terms for which $\alpha > 0.05$ are marked with a double asterisk. Estimates with a single asterisk are significant at the $\alpha = 0.01$ level. Table 12 and Table 13 list the results.

The results are the transformed coefficient estimates. To compute the actual reenlistment rates, use Equation 35, where β is the vector of estimates, and X is the vector of variables observations.

$$P_i = \left[\frac{1}{1 + e^{-\underline{X}\underline{\beta}}} \right] \qquad (36)$$

The variables labels of the tables are as follows:

• Inter	**INTERCEPT**
• Var 1	**BONUS LEVEL**
• Var 2	**REENLISTMENT SYSTEM**
• Var 3	**AFQT SCORE**
• Var 4	**PROMOTION RATE**
• Var 5	**PAY RATE**
• Var 6	**AGE AT ENTRY**
• Var 7	**UNEMPLOYMENT RATE**

UNEMPLOYMENT RATE is not listed on chart. Only two cells include this variable and results are listed here. Cell 52 includes the variable **UNEMPLOYMENT RATE** with a coefficient estimate of 0.105. It is significant at the $\alpha = 0.01$ level. Cell 73 includes the variable **UNEMPLOYMENT RATE** with a coefficient estimate of -0.036. It is significant at the $\alpha = 0.01$ level. The R values are listed under the cell number for each cell.

Table 12. REGRESSION RESULTS BY ZONE

Cell # (R Val)	Inter	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
Cell 1 (0.095)	-0.113 **	0.141		-0.012 *			
Cell 2 (0.000)	-1.141						
Cell 3 (0.260)	-3.422 *	0.141	0.124 *		0.063 *	0.101 *	
Cell 5 (0.085)	-1.154 *		0.102 *	-0.007 *			0.036 *
Cell 6 (0.242)	-2.373 *	0.248 *	0.145 *	-0.009 *	0.065 *	0.093 *	
Cell 7 (0.178)	1.145			-0.033 *			
Cell 8 (0.080)	-0.172 **					0.066 *	
Cell 12 (0.075)	0.581			-0.010			
Cell 15 (0.064)	-0.543 **		0.114 *				0.033
Cell 16 (0.130)	1.001 *			-0.017 *			
Cell 17 (0.135)	2.198 *		0.198 *				-0.084
Cell 18 (0.158)	-0.198 **			-0.011 *	-0.033 *		0.066 *
Cell 22 (0.140)	-1.097	0.209 *		-0.012 *			0.057
Cell 24 (0.10)	0.003 **	0.170 *		-0.009 *			
Cell 26 (0.128)	-1.604 *	0.278 *	0.131 *			0.040 *	
Cell 28 (0.144)	0.940 *	0.179 *		-0.010 *	-0.025 *		
Cell 31 (0.137)	0.646 *	0.200 *		-0.008	-0.029 *		
Cell 37 (0.093)	-0.303 **	0.176	0.160 *				

Table 13. REGRESSION RESULTS BY ZONE (CONTINUED)

Cell # (R Val)	Inter	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6
Cell 38 (0.177)	1.757 *	0.339 *		-0.015 *	-0.025 *		
Cell 39 (0.157)	-0.309 **		0.357 *				
Cell 41 (0.147)	-0.731 *	0.437 *	0.176				
Cell 43 (0.000)	0.464						
Cell 46 (0.120)	-0.681 *	0.142 *	0.252 *		0.017 *		
Cell 47 (0.058)	-2.00 *						0.066
Cell 49 (0.061)	-0.239	0.088	0.175				
Cell 50 (0.179)	1.016 *		0.260 *	-0.021 *			
Cell 51 (0.120)	-0.133 **		0.183 *		0.025 *		
Cell 52 (0.163)	-0.917 *	0.220 *	0.226 *		0.038 *		
Cell 54 (0.120)	0.155 **	0.318		0.012	0.029		
Cell 58 (0.144)	-1.094 *	0.086 *			0.022	0.086 *	
Cell 63 (0.122)	-0.683 *	0.170 *	0.242 *				
Cell 66 (0.100)	-1.967 *	0.188 *	0.149 *			0.046 *	
Cell 70 (0.111)	-0.937 *	0.114	0.168 *	-0.005	0.016		
Cell 72 (0.00)	-0.513 *						
Cell 73 (0.121)	-1.278 *	0.260 *	0.087 *		0.008	0.032 *	
Cell 76 (0.138)	-0.398 *	0.160 *	0.135 *		0.043 *		

LIST OF REFERENCES

1. DeWolf, D. D., Stevens, J. G., and Wood, R. K., "Setting Military Reenlistment Bonuses," paper presented at the Operations Research Society of America, The Institute of Management Science, and the Canadian Operations Research Society joint meeting, Vancouver, BC, 6 May 1989.

2. Center for Naval Analyses Report CRM 87-43, *The Effects of Selective Reenlistment Bonuses on Retention,* by D. J. Cymrot, March 1987.

3. The RAND Corporation Report R-1502-ARPA, *Effect of the Variable Reenlistment Bonus on Reenlistment Rates: Empirical Results for FY1971,* by J. H. Enns, 1975.

4. Lakhani, H., and Gilroy, C. L., "Army Reenlistment and Extension Decisions by Occupation," in *Army Manpower Economics,* pp. 225-256, Westview Press, 1985.

5. The RAND Corporation Report R-3348-FMP, *Using Survivor Functions to Estimate Occupation-Specific Bonus Effects,* by D. F. Kohler, March 1988.

6. The RAND Corporation Report R-3199-MIL, *Reenlistment Bonuses and Retention Behavior,* by J. R. Hosek and C. E. Peterson, May 1985.

7. U. S. Army Concepts Analysis Agency Report CAA-SR-82-6, *Selective Reenlistment Bonus Study,* by H. J. Lenius and others, August 1982.

8. The RAND Corporation Report R-2468-MRAL, *Models of the First-Term Reenlistment Decision,* by W. K. Chow and J. M. Polich, September 1980.

9. The RAND Corporation Report R-2884-MRAL, *Analysis of Second-Term Reenlistment Behavior,* by J. R. Hiller, September 1982.

10. Higham, R. P., *A Multiple Linear Regression Model for Predicting Zone A Retention by Military Occupational Specialty*, Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1986.

11. Daula, T. V., and Baldwin, R. H., "Reenlistment Decision Model: Implications for Policy Making, " in *Army Manpower Economics*, pp. 203-221, Westview Press, 1985.

12. Systems Research and Applications Corporation Report 85-0058, "Dynamic Model of Military Reenlistment Behavior," by Black, M., and others, June 1987.

13. The RAND Corporation Report P-7007, *Active and Reserve Force Attrition and Retention: A Selected Review of Research and Methods*, by Z. D. Doering and D. W. Grissmer, March 1985.

14. Defense Manpower Data Center Report TR-84-3, *Why Service Members Leave the Military: Review of the Literature and Analysis*, by D. Boesel and K. Johnson, April 1984.

15. The RAND Corporation Report R-2152-ARPA, *First-Term Reenlistment Intentions of Avionics Technicians: A Quantitative Analysis*, by W. D. Perry, October 1977.

16. The RAND Corporation Report R-3117-MIL, *The Retention of High-Quality Personnel in the U. S. Armed Forces*, by M. P. Ward and H. W. Tan, February 1985.

17. Warner, J. B., *Analysis of Intelligence and Academic Scores as a Predictor of Promotion Rate for U. S. Army Noncommissioned Officers*, Master's Thesis, Naval Postgraduate School, Monterey, CA, June 1987.

18. Navy Personnel Research and Development Center Report TR 81-2, *Job Satisfaction Measures as Predictors of Retention for Navy Enlisted Personnel*, by M. H. Royle and D. W. Robertson, December 1980.

19. Navy Personnel Research and Development Center Report SR 81-25, *The Effects of Unmet Expectations, Satisfaction, and Commitment on the Reenlistment Intentions of First-Term Enlisted Personnel*, by A. H. Farkas, August 1981.

20. Center of Naval Analyses Report CRC 454, *Relating Attitudes Towards Navy Life to Reenlistment Decisions*, by J. W. Fletcher and K. L. Giesler, September 1981.

21. *SAS User's Guide: Basic*, Version 5 ed., SAS Institute Inc., 1985.

22. Chambers, J. M., and others, *Graphical Methods for Data Analysis*, Wadsworth, Inc., 1983.

23. Freeman, D. H., and Jekel, J. F., "Table Selection and Log-Linear Models," *Journal of Chronic Diseases*, v. 33, pp. 513-524, 1980.

24. Kotze, T. J., "The Log-Linear Model and Its Application to Multi-Way Contingency Tables," in *Topics in Applied Multivariate Analysis*, Cambridge University Press, 1982.

25. Grizzel, J. E., and others, "Analysis of Categorical Data by Linear Models," *Biometrics*, v. 56, pp. 489-504, September 1969.

26. Bishop, Y. M., and others, *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, 1975.

27. Conover, W. J., *Practical Nonparametric Statistics*, 2d ed., John Wiley & Sons, Inc., 1980.

28. Bishop, Y. M., "Effects of Collapsing Multidimensional Contingency Tables," *Biometrics*, v. 27, pp. 545-562, September 1971.

29. Whittemore, A. S., "Collapsibility of Multidimensional Contingency Tables," *Journal of the Royal Statistical Society*, v. 40, no. 3, pp. 328-340, 1978.

30. Larsen, R. W., *The Aggregation of Population Groups to Improve the Predictability of Marine Corps Officer Attrition Estimation,* Master's Thesis, Naval Postgraduate School, Monterey, CA, December 1987.

31. Harrell, F. E., "The LOGIST Procedure," *SUGI Supplemental Library User's Guide,* v. 5, pp. 181-202, 1986.

32. Mosteller, F., and Tukey, J. W., *Data Analysis and Regression, a Second Course in Statistics,* Addison-Wesley Publishing Company, 1977.

33. Misiewicz, J. R., *Extension of Aggregation and Shrinkage Techniques Used in Estimation of Marine Corps Officer Attrition Rates,* Master's Thesis, Naval Postgraduate School, Monterey, CA, September 1989.

34. Box, G. E., and Tiao, G. C., "Intervention Analysis With Application to Economic and Environmental Problems," *Journal of the American Statistical Association,* v. 70, pp. 70-83, May 1975.

35. Judge, G. G., and others, *The Theory and Practice of Econometrics* , 2d ed., John Wiley & Sons Inc., 1985.

36. Center for Naval Analyses Report CRC 476, *Determinants of Navy Reenlistment and Extension Rates,* by M. S. Goldberg and J. T. Warner, December 1982.

37. Center for Naval Analyses Report CRM 87-91, *Extension and Subsequent Reenlistment in the United States Marine Corps* by D. J. Cymrot, June 1987.

38. Center for Naval Analyses Report CRM 88-45, *Defining the Population Making Reenlistment Decisions* by D. J. Cymrot, June 1988.

39. Center for Naval Analyses Report CRM 85-38, *Retention Costs of the GI Bill and the Draft: New Evidence From the Navy's Enlisted Force,* by J. S. Thomason, January 1985.

40. Cohen, L., and Reedy, D. E., "The Sensitivity of First Term Navy Reenlistment to Changes in Unemployment and Relative Wages," *Naval Research Logistical Quarterly*, v. 26, n. 4, pp. 695-709, December 1979.

41. The RAND Corporation Report R-3211-MIL, *Voluntary Terminations from Military Service*, by R. M. Solzenberg and J. D. Winkler, May 1983.

42. Brown, C., "Military Enlistments: What Can We Learn from Geographic Variations?" *American Economic Review*, v. 75, n. 1, pp. 228-234, March 1985.

43. U. S. Department of Labor, Bureau of Labor Statistics, *Employment and Earnings*, Government Printing Office, Washington, DC, July 1989.

44. U. S. Army Research Institute for the Behavioral and Social Sciences Report TR797, *A Multivariate Analysis of Determinants of Reenlistment: A Decision-Making Model for Enlisted Personnel*, by A. L. Smith Jr., June 1988.

45. U. S. Department of Defense, Office of the Assistant Secretary of Defense, Manpower, Installations and Logistics, DoD 1312.1-M, *Occupation Conversion Manual*, Government Printing Office, Washington, DC, September 1984.

46. The RAND Corporation Report R-1565-ARPA, *An Imputation Method for Estimating Civilian Opportunities Available to Military Enlisted Men*, by A. P. Massell, July 1975.

47. U. S. Department of Labor, Bureau of Labor Statistics, *CPI Detailed Report*, Government Printing Office, Washington, DC, May 1989.

48. U. S. Department of Commerce, Bureau of Economic Analysis, *Survey of Current Business*, Government Printing Office, Washington, DC, March 1989.

49. *SAS User's Guide: Statistics*, Version 5 ed., SAS Institute Inc., 1985.

50. Draper, N. R., and Smith, N., *Applied Regression Analysis*, 2d., John Wiley & Sons, Inc., 1981.

51. The RAND Corporation Report R-1306-EDA NIH, *Univariate and Multivariate Log-Linear and Logistic Models*, by M. Nerlove and S. J. Press, December 1973.

52. Larsen, R. J., and Marx, M. L., *An Introduction to Mathematical Statistics and Its Applications*, Prentice-Hall, 1986.

53. Harrell, F. E., "The LOGIST Procedure," *SUGI Supplemental Library User's Guide*, v. 5, pp. 181-202, 1986.

54. Efron, B., "Bootstrap Methods: Another Look at the Jackknife," *Annuals of Statistics*, v. 7, pp. 1-26, 1979.

55. Freedman, D. A., "A Note on Screaning Regression Equations," *The American Statistician*, v. 37, pp. 152-155, May 1983.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Technical Information Center Cameron Station Alexandria, VA 22304-6145	2
2.	Library, Code 0142 Naval Postgraduate School Monterey, CA 93943-5002	2
3.	Commander U. S. Total Army Personnel Command ATTN: TAPC-EPT-B 2461 Eisenhower Avenue Alexandria, VA 22331-0457	5
4.	Headquarters, Department of the Army ATTN: DAPE-MBB-P (Room 2D669) Washington, DC 20310	5
5.	Commander U. S. Army Research Institute for the Behavioral and Social Sciences ATTN: PERI-RP (Robert Tinney) 5001 Eisenhower Avenue Alexandria, VA 22333	1
6.	Commander U. S. Army TRADOC Analysis Command ATTN: ATRC Fort Leavenworth, KS 66027-5200	2
7.	Deputy Undersecretary of the Army for Operations Research Room 2E261, Pentagon Washington, DC 20310	2
8.	Professor Laura D. Johnson Naval Postgraduate School, Code 55Jo Monterey, CA 93943-5000	2
9.	Professor Donald P. Gaver Jr Naval Postgraduate School, Code 55Ga Monterey, CA 93943-5000	2
10.	Department Chairman Department of Operations Research Naval Postgraduate School, Code 55 Monterey, CA 93943-5000	1

11. Captain James G. Stevens 1
Naval Postgraduate School, Code 55
Monterey, CA 93943-5000

12. Director 2
Defense Manpower Data Center
ATTN: Lynn Routsong
99 Pacific Street, Suite 15A
Monterey, CA 93940-2453

13. Helen Davis 1
Users Services Group
W. R. Church Computer Center
Naval Postgraduate School, Code 0141
Monterey, CA 93943-5000

14. Captain Michael J. Streff 2
12068 Stallion Court
Woodbridge, VA 22192

6/9-588

Thesis
S8244 Streff
c.1 The estimation of
United States Army
reenlistment rates.

AMERICAN BINDERY
TOPEKA KANSAS

The estimation of United States Army ree
3 2768 000 86733 7
DUDLEY KNOX LIBRARY